Overview

This study aims to develop a methodology to classify the lying behaviour of dairy cows using noisy positioning data. This is achieved by preprocessing the data then applying changepoint analysis and bagged decision trees. The authors successfully classify lying behaviour to a high accuracy, which has not been done using spatial positioning data in an automated non-invasive way to date. This could help refine farm management to improve dairy cow health and welfare. Additional details and justification in the methodology would strengthen this manuscript, such as outlining the amount of data lost after each pre-processing step and demonstrating the minimal change in accuracy with different splits in the data between the training and test sets.

AU: Thank you for your valuable comments and suggestions. We have addressed them below and believe this has improved the manuscript, hopefully to your satisfaction.

Major

1. Section 3.3. The authors have clearly outlined the data editing steps. Outlining the amount of data lost during each data editing step would be useful, to show the reader specifically how much data was excluded within the average of 43% lost per day (including 'correct' data during milking). The use of a flow chart may help. Furthermore, it would be beneficial if the details of the data exploration, which the management of missing data was based on, were included (L182-184), perhaps as Supplementary Material).

AU: When expressed on a "number of seconds no data was available" per day, on average 43% of the data were missing (i.e., on average for 37.152s of the 86400s/day there was no measurement available). We did not *exclude* data during the data editing process, only did we impute some of these missing values using the data imputation techniques as described in L177 to 192. As this seems unclear in the manuscript, we rewrote this in L178 to 180. Furthermore, data exploration is a non-linear process, that we believe has little value in detailing. The repository with the code contains multiple scripts that do this, and we hope this is sufficient for interested readers to explore. We made a reference to these codes in L158-159

We added some figures to support the data exploration to appendix as well and referred to them in text L168, and 173.

- A figure giving the missing data per cow;
- A figure with an example of data outside the barn (when a tag was excluded because it was not attached to a cow anymore).

2. L177-179. It is not clear whether the data points either side of these instances were always close to the edge value. If not, why were these instances replaced by the edge value rather than excluded, as was done with successive data points out of the barn edges?

AU: In the decision process how to deal with measurement noise and missing values, we preferred to exclude as little data points as possible (and even chose to make data imputations were sensible). As pointed out in the text, there is a big difference between the situation when a tag was presumably still attached to the cow (few data outside the barn edges, measurement error) or when a tag/collar was lost by a cow and placed outside the barn by the herdsmen (long series of measurements outside the barn edges). In the first case, first deleting single measurements to later impute the data would have the same effect as replacing the measurement by the edges in one step, so this was preferred. In the second case, no data on actual cow positions are collected in the first place, so no editing step can help solving this issue. As this is a hardware problem and no data processing or accuracy issue, how often this happens really depends on the quality of collars etc., which is not in the scope of this study to evaluate.

3. Section 3.4. The authors explain that the (z)-position was found to be the most unreliable and noisy of all three coordinates (L213-215). Is it not clear why the (z)- coordinate was therefore used in the classification at all, apart from because it is straightforward (L226-227). Would a

measure derived from the coordinates have been useful? e.g., VeDBA and SCAY (see Vázquez Diosdado et al., 2015).

AU: Indeed, the hardware specifics of the UWB system induced a large error on the z-coordinates. Despite this, the data exploration (see also added Figure A3 in supplementary = distributions of the z-values) still clearly showed that in many cases, "on average" the z-values contained the right information (i.e.. height of the cow), but that more advanced data processing was needed to unlock this information for detecting lying behaviour. As including the z-coordinate in the changepoint analysis comes at little extra (computational) cost, there was no reason not to consider it. Accelerations were, for this sensors, not available. For x and y, we indeed use a derived metric (centerdistance) instead of the raw measurements, as this also makes sense "biologically" when considering the target behaviour. For using the z-coordinate, we could not come up with a metric that would be meaningful in this regard. We included some discussion that inclusion of more derived metrics such as VeDBA and SCAY can be a very valuable addition when different data (sources) are available in L415 to 417.

4. **Section 3.5**. The data-split approach could do with further justification. Was the use of a validation dataset or K-fold cross validation to increase robustness considered?

AU: As explained in section 3.6, a cross-validation has been used to identify the optimal set of hyperparameters for the classification algorithm (section 4.4). For model evaluation, however, we could indeed assess different data splits and how this influenced performance (see also answer to point 5.).

5. **L267-269**. Please provide results to demonstrate increasing the data used for training did not increase the accuracy. It would be useful to see the results of different combinations.

AU: We implemented the classification models using 3, 4, 5, 6, 7, 8 and 9 days of data, and indeed saw that only limited improvements in the accuracy of the classification were reached with increasing the number of days in the training data. Therefore, and considering (1) complexity and readability of the manuscript; (2) the argumentation that using as little as possible training data is crucial for potential on-farm implementations, we did not add the extra results to the manuscript. We added a line to the manuscript to further explain this (L258-259).

Table 1. Classification results when 3 to 8 days of data are used for training the classifier.

days in training	No. 'lying' segments	No. correctly classified as 'lying'	% correct	No. 'non-lying' segments	No. correctly classified as 'non-lying'	% correct
3	8286	7614	91.8	6602	6105	92.5
4	7281	6722	92.3	5911	5438	92.0
5	6264	5804	92.7	5169	4738	91.7
6	5625	5242	93.2	4599	4234	92.1
7	4738	4434	93.6	3858	3540	91.8
8	3846	3610	93.9	3050	2794	91.6

6. **L271-277**. The division between training and test data is set to be different for the cow identify-based data split and the time-based data split, but why this was done is unclear. In Section 4.4, please state the final division as a percentage, as in Section 4.3 (35% training).

AU: we added that for the split based on time, 25.66% of the data are in the training set (5138 segments) and 74.33% are in the test set (14888) segments in L264-266.

7. **L286-288**. It is explained that this classification algorithm performed best- is this in relation to other algorithms and can you provide evidence?

AU: this was part of an initial exploration step, in which different classification algorithms were tested and compared for all the data and subsets thereof. While we agree that a comparison of the performance can be interesting, there's very little novelty in it (application of machine learning for classification tasks), and therefore it was excluded from the manuscript. We added some lines in the text L277 to 279 to explain this, and make the nuance that although the techniques applied here can be generalised to different behaviour classification problems, which algorithm performs best is probably dependent on the data, the task etc..

Minor

8. L50. Good explanation- please briefly expand on how they are much less homogenous.

AU: We added that cows often are kept in production systems where cows from different parities, ages, lactation stages are mixed L48-49.

9. L57-61. Please add references here e.g., Boyland et al. (2015) and Chopra et al. (2020).

AU: We added the references you suggest L58-59.

10. L57-81. The authors could mention other indoor positioning systems that have been used to monitor behaviour over time, such as local positioning systems (LPS), and compare these to uwb-based positioning systems.

AU: We added L59-61 and L80-81 information and a reference to an article (https://www.sciencedirect.com/science/article/pii/S1537511006004053) that compares and discusses indoor positioning technologies for cows. The main advantage of UWB are its wide frequency spectrum, it battery usage and the level of accuracy it can provide compared to e.g. GPS based systems.

11. L72-73. Add references(s) showing findings of changing lying behaviour in relation to lameness e.g., Barker et al. (2018).

AU: We added references to Barker and Weigele (https://www.journalofdairyscience.org/article/S0022-0302(17)31201-8/pdf) who discuss the effect of (moderate) lameness on dairy cow lying behaviour L71.

12. L93-95. Has research on uwb with a view for livestock applications been conducted previously? If so, please provide a brief overview and references.

AU: There has been quite many work conducted on developing applications of UWB for livestock, as proven by the many companies that sell these sort of technologies. Unfortunately, we couldn't find a scientific reference that compares the different systems/applications.

13. L99-103. 'previous research' regarding these limitations needs a citation.

AU: we added citations to Ren et al. (2022) in L99.

14. L127-131. Please give the accuracy of the uwb-positioning tags and the accelerometers.

AU: We added the accuracy of the uwb in both directions L124-128.

15. L133. Please check- is the latter meant to read (y)-position?

AU: it's indeed an error, and it should read (z) position. We corrected it in the manuscript.

16. L171-173. Were there not any instances when cows were lying down in the feeding area? Was the data filtered to y > 11.5m? This is unclear.

AU: We checked and for the current dataset, there were no recorded lying bouts in the slatted flooring area. We indeed filter the data to exclude measurements in this area, as clarified L165

17. L202-203. Perhaps include why this was done i.e., to reduce noise.

AU: We added this in the manuscript L194-195.

18. Figure 1. The visualisation of the barn layouts is useful. Please move the figure so it is directly under Section 3.3, and please move the figure legend so it is directly under the figure. It may help to stick to one unit or measurement in the main text and figure e.g., m (see L166-173). In the main text, the minimum and maximum x and y coordinate values do not match those in the figure (see L166-173). Please ensure you are consistent with which sides of the barn layouts were considered x and y too.

AU: We implemented your suggestions and remade the figure 1 for clarification. Indeed, the first figure was an older version with some errors in it.

19. The positioning of most figures and tables needs improving- please place directly under associated text. E.g., I suggest moving Fig 3-5 to the results section. Please also state when a table/figure is in the Appendix.

AU: we improved positioning of the figures to appear under the relevant sections

20. Figures and tables are often not interpretable independent to the main text- this could be improved by providing additional details in the figure/table legends.

AU: we extended to captions to be more informative

Reviewed by John Fredy Ramirez Agudelo, 03 May 2022 08:35

This is a very interesting work in which the authors explore the use of position data to predict the liying behavior.To do this, the authors suggest the analysis of data recorded by ultra-wide band positioning tags on the upside of a neck collar. This methodology may represent an advantage over other methods established for this purpose (accelerometers) or even emerging ones (computer vision).

AU: Thank you for your kind words.

Below are some comments that can make the paper easier to read, especially for those readers unfamiliar with this type of sensors and data analysis.

AU: Thank you for your suggestions. We hope we have improved the manuscript to your satisfaction.

L20. To avoid misinterpretation, you can add the words "barn" in: "distance from the center"

AU: Done

L27-31. Although the performance of the prediction is presented, it is also interesting to briefly present, for example, the average difference between the liying time registered by the accelerometers and by the new methodology; or the average number of undetected lying bouts per day per cow

AU: we added the information on lying duration/day to the abstract. However, number of undetected lying bouts was not calculated, as we did not assign each segment to a lying bout with a different ID separately.

L72-73. Could you add some references that support the statement made in these lines?, like in L75

AU: we added references to Barker et al. and Weigele et al. as also requested by reviewer 1 L71.

L91-93. Could you add a reference where the reader can find information about video-based systems?

AU: We added references to McDonagh et al. (2021) L89.

L116-120. The methodological aspects described in these lines are addressed later, so these lines can be ignored.

AU: We deleted these lines

L131-135. A technical aspect is pointed out, but the implications are not clear. If it's just a manufacturing feature, it can be ignored.

AU: As it determines the fact we need to deal with different levels of inaccuracies, this is important information. We added the implications to the manuscript L127-128.

L165-166. Given the importance of the pre-specified origin (x,y,z)=(0,0,0), could you provide more information on the aspects that were considered when selecting it? Was there a single point of origin for both barns? Could the results be improved by using more reference points, mainly for (z)-position ?

AU: We added some more details on how important this origin is. It is mainly a hardware technicity, so we could not change it, and as it's an intrinsic of the sensor system, adding more references is not possible. Details are added in L154-155.

L166-169. The values of the x and y axes in Figure 1 do not correspond to the description of these lines.

AU: Indeed, this was not the final plan, and we changed the figure and its units for clarification to agree with the correct situation.

L171-172 Figure 1 do not correspond to the description of these lines. "When the y was larger than 11.5m, the animals were in the slatted flooring (feeding) area."

AU: indeed, this was not the final plan, and we changed the figure and its units for clarification to agree with the correct situation.

L182. It is not clear what it means: "These measurements were replaced by missing values."

AU: it means that the 'time stamp' was retained in the data, but the x,y and z values were deleted and replace with "NaN = Not a Number". We clarified this L174-175.

L267-269. Did you perform preliminary tests to ensure this? Is there a minimum number of training segments that ensures optimal performance of the methodology?

AU: We did multiple tests, as also responded to the remarks of Reviewer 1. Hopefully the answer there is to your satisfaction. We also added the information in L257-258

L271. This was already mentioned: (alike the more classical machine learning approach)

AU: we deleted the repetition.

L286-292. Could you add some references that support the statements made in these lines?

AU: we added references to Dietterich, Breiman (https://dl.acm.org/doi/pdf/10.1145/212094.212114) (Breiman, L. (1996). Bagging predictors. Machine Learning 24, 123-140) in L274-276.

Below Table 2, information about the variables and values presented is required.

AU: We added the information the clarify the tables.

L342-344. But there is essentially no evidence to expect this.

AU: We nuanced this sentence to explain that there is indeed no evidence to verify or expect this, but that no sensor is faultless L331-332.

In Figure 2 and 4. Replace the word "Class" ----> "ground truth" or "Truth". Also, information about the values presented is required.

AU: we adjusted the figures to clarify them.