Response to Reviewer 3 Comments

Manuscript ID: sensors-4553923

Title: Handcrafted Features versus Learned Representations for IMU-Based Goat Behavior Classification: Accuracy, Validation Design, and Computational Cost

1. Summary

Thank you very much for taking the time to review this manuscript. We are particularly grateful for the detailed reading and the annotated manuscript. Please find the detailed responses below and the corresponding revisions marked in blue in the resubmitted manuscript. Each response states briefly what was changed and where, followed by the revised text; page and line numbers refer to the revised manuscript. Before the point-by-point responses, we report three corrections that we identified during the revision (Section 2); none of them changes the conclusions.

2. Corrections identified by the authors

First, while re-examining the data pipeline for this revision, we found that the integer codes of the two focal classes had been interchanged in the curated dataset relative to the class names used in the manuscript: the windows reported as Rumination were in fact eating, and vice versa. We confirmed this against the original annotation files and their metadata, by byte-level matching of raw signal windows between the manually annotated set and the curated dataset, and by the motion signature of each class (gyroscope magnitude about five times higher in the windows now labeled Eating than in those labeled Rumination). Because Macro F1 is invariant to a permutation of class names, no Macro F1 value, model ranking, contrast, sampling-rate or latency result changes. We have corrected Table 1, the per-class columns of Tables 5 and 6 and of Table S2, Figure 6 and every sentence that names a class, and have rewritten the two interpretations that depended on the class names: the ordering of class difficulty and the class-dependent deficit of Model F:

> "Eating in a pen-fed system involves large, sustained head and neck movements toward the feed trough, which produce a high-amplitude, readily separable signature at the collar. Rumination is a low-motion state in which the only consistent signal is the small rhythmic jaw movement of chewing [37]; against this quiet background it shares much of its signal envelope with resting and other inactive states." (**Section 4.1, page 20, lines 692–697**)
>
> "The deficit was strongly class-dependent, smallest for Eating (0.040 under LOSO), intermediate for Rumination (0.123), and largest for Other (0.218), consistent with the class-difficulty ordering. One plausible explanation is that the mean-squared reconstruction objective is dominated by high-amplitude signal components. The large head movements of eating contribute substantially to reconstruction error if omitted, so a representation optimized for reconstruction has reason to retain them. Separating rumination from the other low-motion states, by contrast, depends on small-amplitude chewing pulses [37] that contribute little to reconstruction fidelity and are therefore not prioritized by the objective." (**Section 4.2, page 20, lines 715–723**)

Second, per-subject Macro F1 under LOSO is now averaged over the classes present in each subject's held-out partition. Seven of the 15 LOSO subjects contributed no Other windows (Table S4), and the previous computation had assigned them an F1 of zero for that class. The per-subject values in Table 4 and the statistics derived from them, such as the ratio of between-subject variability to the spread among models (now roughly five-fold rather than 4.5-fold), have been recomputed; aggregate Macro F1 is unaffected. We also found that the largest difference in aggregate Macro F1 between the re-executed confusion matrices of Figure 6 and Table 5 is 0.003, not 0.005 as previously stated, so run-to-run variation is now described as more than half, rather than larger than, the 0.0048 difference between Models A and B:

> "…the mean ± SD of per-subject Macro F1 (n = 15), where each subject's Macro F1 is averaged over the classes present in its held-out partition." (**Section 2.7, page 12, lines 424–426**)
>
> "The full range of per-subject mean Macro F1 across Models A–E was 0.025, smaller than the between-subject standard deviation by a factor of approximately 5." (**Section 3.1, page 14, lines 527–529**)
>
> "…per-class F1 derived from them differs from Table 5 by at most 0.006, and aggregate Macro F1 by at most 0.003, reflecting non-deterministic GPU execution." (**Section 3.2.1, pages 14–15, lines 548–550**)
>
> "Re-execution under identical seeds shifted the aggregate Macro F1 of Models A, B, and D by up to 0.003 (Section 3.2.1), more than half the 0.0048 difference between Models A and B under LOSO; the smallest between-model gaps are therefore of the same order as run-to-run variation alone." (**Section 4.3, page 21, lines 767–770**)

Third, in preparing the new Figure 1, which now shows the sensor axes relative to the animal's body, we found that the axis assignment in the original Section 2.1 was incorrect. The text stated that gravity was registered on the −Y axis and took Y as the dorsoventral axis; gravity is in fact carried by the X axis, which runs along the collar strap. No result depends on this description, because all models received the raw axis-native signals. The corrected description reads:

> "The X axis runs along the collar strap (positive ventrally), the Y axis is normal to the neck surface (positive outward), and the Z axis runs along the neck (positive cranially)." (**Section 2.1, page 4, lines 136–137**)

We apologize for these errors. We have also re-checked every citation against its source and revised the wording where it did not match the cited study; these changes are marked in blue as well.

3. Point-by-point response to Comments and Suggestions for Authors

Responses to the general comments are given first (Comments 1–7), followed by the comments in the annotated PDF (Comments 8–57). For the annotations, the page of the originally submitted PDF and the highlighted text are given in parentheses.

3.1 General comments

Comments 1: I recommend that the authors expand their literature search to include relevant literature, particularly on beef bulls, cows, and calves, as well as papers on dairy cows. Studies have implemented or tested approaches similar to those described by the authors. Including them and discussing their findings alongside yours would strengthen the paper and properly place the novelty of your study within the field.

Response 1: Thank you. We agree. We have added five studies in cattle, namely Rahman et al. (2018), Camargo et al. (2024; beef bulls), Riaboff et al. (2020; dairy cows), Vázquez Diosdado et al. (2015; housed dairy cows) and Carslake et al. (2021; dairy calves), and one in goats, Kamminga et al. (2018; collar-mounted sensors). They are cited in the Introduction (paragraphs 1, 2 and 4), for the methodological choices in Sections 2.1–2.3 and in the Discussion (Sections 4.3 and 4.5), for example:

> "In ruminants, rumination and eating time are among the most informative behavioral indicators for health surveillance in dairy systems and for health and reproductive monitoring in beef cattle [2,3]." (**Introduction, page 2, lines 54–56**)
>
> "Statistical and spectral features are extracted from signal windows and, often after feature selection, classified with decision trees or tree ensembles such as random forests, AdaBoost, or gradient boosting [7-10]." (**Introduction, page 2, lines 67–69**)
>
> "All downstream models were trained and evaluated on raw axis-native signals, which places the burden of orientation invariance on the model and not on a preprocessing pipeline [19]." (**Section 2.1, page 4, lines 140–142**)

The other passages citing these studies are quoted in Responses 15 (Introduction, paragraph 4), 18 (Section 4.3, paragraph 4), 22 (Sections 2.2 and 4.5), 23 and 24 (Section 2.3) and 50 (Section 4.3).

Comments 2: The authors used a mixed labelling approach (computer-vision and visual observation) for ground truth. A concern is that the company that provided the sensors also provided the computer-vision software. The authors should address this and discuss whether it could introduce data leakage or compounding uncertainty that affects their results.

Response 2: Thank you for raising this concern. We should first clarify that the automatic labeler does not use computer vision: it operates on the IMU signal, and the BoditToolBox software is only an interval-marking tool used by the observers, who worked from the video alone:

> "Five trained observers, each responsible for a distinct subset of animals with no overlap, first annotated rumination and eating episodes from the reference video using the BoditToolBox video annotation software (version 1.0, Bodit Co., Ltd., Republic of Korea), an interval-marking tool; the IMU waveform was not shown to the observers." (**Section 2.2, page 4, lines 153–157**)

The leakage concern therefore applies to the IMU-based labeler. Section 2.2 now describes its stages and the share of final windows from manual annotation and from the labeler, the video check of 600 retained windows is now reported with Wilson 95% confidence intervals, and Section 4.5 states that the circularity introduced by the labeler's KNN stage is most relevant to Model A:

> "A two-stage automatic labeler was then applied to the full recordings (1,416,961 windows over 164 animal-days). … The binary screening model was developed by the sensor manufacturer, and its architecture is not disclosed here. Because the KNN stage uses the same handcrafted feature family as Model A, the labels are not fully independent of Model A's input representation. Of the 132,887 windows in the final dataset, 78,320 (58.9%), including Other windows, originated directly from manual annotation and 54,567 (41.1%) from the automatic labeler." (**Section 2.2, pages 4–5, lines 158–172**)
>
> "Agreement between the assigned labels and the human video assessment was 99.3% overall (596/600; Wilson 95% CI 98.3–99.7%): Rumination 98.5% (95.7–99.5%), Eating 99.5% (97.2–99.9%), and Other 100.0% (98.1–100.0%)." (**Section 2.2, page 5, lines 175–177**)
>
> "About 41% of the final windows were assigned by the automatic labeler, whose KNN stage uses the same handcrafted feature family as Model A; this potential circularity is therefore most relevant to Model A, and if anything biases the comparison against the learned representations that matched it." (**Section 4.5, page 23, lines 841–844**)

The architecture of the manufacturer's binary screening model is not disclosed, as stated in Section 2.2 (quoted above) and in the Limitations, and the manufacturer's role is stated, as in the original submission, in the Conflicts of Interest:

> "The binary screening stage of the labeler was supplied by the sensor manufacturer and its architecture is not disclosed, so the reliability of that stage is characterized empirically, by agreement with synchronized video, rather than by specification." (**Section 4.5, page 23, lines 877–880**)
>
> "Bodit Co., Ltd. provided the sensor hardware and the labeling pipeline and contributed to data acquisition. The comparison of model architectures reported here, the interpretation of the results, and the decision to publish were determined by the academic authors." (**Conflicts of Interest, page 25, lines 951–954**)

Comments 3: No inter-rater reliability is presented to assess the consistency of the labelling over time. The authors should also better describe the observations and the composition of the subsampled data.

Response 3: Thank you. Each animal was annotated by one of five observers with no overlap, so inter-rater reliability could not be computed, and intra-observer consistency over time was not tested either. The first point is now stated in the Methods, and both are stated as a limitation:

> "Because each animal was annotated by a single observer, inter-rater reliability could not be computed." (**Section 2.2, page 5, lines 177–178**)

> "Inter-rater reliability of the manual annotation could not be assessed because each animal was annotated by a single observer, and intra-observer consistency over time was not tested." (**Section 4.5, page 23, lines 847–849**)

The only re-examination of the labels against the video was the check of 600 retained windows (Section 2.2, paragraph 3; quoted in Response 2), which agreed with the assigned labels in 99.3% of cases; it indicates the accuracy of the retained labels rather than the stability of individual observers.

The observations are now described in more detail, and Section 2.3 explains the origin of the Other class and the subsampling:

> "Five trained observers, each responsible for a distinct subset of animals with no overlap, first annotated rumination and eating episodes from the reference video … This manual stage covered 14 animals over 32 animal-days and yielded 43,119 labeled rumination and eating windows." (**Section 2.2, page 4, lines 153–158**)
>
> "It was drawn mainly from the subset of (subject, day) recordings for which annotation covered the full 24-hour period … Seven fully annotated animal-days from six animals contributed the large majority of Other windows, with the remainder drawn from unlabeled intervals of partially annotated animal-days. … Windows of all three classes were then randomly subsampled without replacement (NumPy RandomState, seed = 42) to a comparable per-class count, approximately 44,000 each, giving an aggregate ratio of about 1 : 1 : 1." (**Section 2.3, page 5, lines 201–211**)

The number of animals and days is now given in the caption of Table 1 (quoted in Response 27), and the number of windows per subject and class in the new Table S4.

Comments 4: I suggest the authors emphasize the LOSO results rather than the K-fold results, even though the differences are not as significant. LOSO is a significantly more robust validation method and produces more realistic results for this type of data. I suggest the authors refer to dairy cattle studies (some of which were cited by the author, such as Riaboff et al.) to incorporate statistical tests to compare models. GLM models would be more appropriate for many of the comparisons.

Response 4: Thank you. We agree. LOSO is now the primary result: it is reported first in the Abstract, in Tables 4, 5 and 7 and in Figures 5 and 7, Table 9 and Figure 9 have been recomputed on LOSO Macro F1 (quoted in Response 45), and Section 2.7 states why:

> "…under leave-one-subject-out (LOSO) and stratified 5-fold cross-validation (CV). A 24-dimensional learned representation matched a 250-dimensional handcrafted feature set (Macro F1 0.810 versus 0.814 under LOSO, 0.876 versus 0.876 under 5-fold CV)…" (**Abstract, page 1, lines 35–38**)
>
> "Figure 5. The two validation schemes. (a) Subject-level leave-one-subject-out over the 15 subjects with at least 50 windows per target class, the primary scheme; (b) Stratified 5-fold cross-validation…" (**Figure 5, page 12, lines 429–430**)

> "Figure 7. Macro F1 as a function of sampling rate for all six models under (a) subject-level LOSO and (b) stratified 5-fold cross-validation." (**Figure 7, page 17, lines 616–617**)
>
> "As the appropriate scheme for assessing predictive performance on new animals, LOSO is treated as the primary result." (**Section 2.7, page 11, lines 410–411**)

Cattle studies that compared validation designs are now cited in the Introduction and in Section 4.3 (see Response 50). For the model comparison we fitted a linear mixed model, the extension of the general linear model to repeated measures, with subject as a random intercept, because all models were evaluated on the same 15 subjects:

> "A linear mixed model was fitted with model (A–F) as a fixed effect and subject as a random intercept, using R version 4.6.1 [31] and the lme4 package [32]." (**new Section 2.11, page 13, lines 491–492**)
>
> "The mixed model indicated a significant overall effect of model ($F(5, 70) = 13.1$, $p < 0.001$), but no pairwise contrast among Models A, B, D, and E reached significance (Tukey-adjusted $p \geq 0.97$)." (**Section 3.1, page 13, lines 503–505**)

Comments 5: I recommend the authors use computational costs rather than deployment costs. Also, add the programming language and packages throughout the manuscript.

Response 5: Agree. The title now ends "Accuracy, Validation Design, and Computational Cost", and "computational cost" is used wherever the text refers to what was measured. Package and function names are now given at the main steps of the Methods: for example, NumPy RandomState in Section 2.3 (quoted in Response 3), scikit-learn StratifiedKFold in Section 2.7 and R with lme4 and emmeans in Section 2.11 (quoted in Response 35). Section 2.10, retitled "Reproducibility and Software Environment", cites each package and gives the versions of Python, PyTorch, LightGBM and scikit-learn:

> "Stratified 5-fold cross-validation (5-fold CV) partitions the full dataset into five folds while preserving the class distribution in each (scikit-learn StratifiedKFold, shuffle = True, random_state = 42)." (**Section 2.7, page 11, lines 412–414**)
>
> "All models were trained and evaluated in an identical software environment: Python 3.12.13, PyTorch 2.11.0 [28] with CUDA 13.0, LightGBM 4.6.0 [21], scikit-learn 1.8.0 [29], NumPy [30], and SciPy [27]." (**Section 2.10, page 13, lines 480–482**)

Comments 6: I also recommend that the authors reflect on how their findings advance knowledge in the field. The discussion includes limitations but could be strengthened.

Response 6: Thank you. We agree. A paragraph at the end of the Limitations now states what the study contributes relative to the existing literature and what its practical results are:

> "Within these limits, the study contributes three things that the existing literature on ruminant behavior classification rarely provides together: a controlled comparison in which feature source, classifier, and training objective are varied one at a time on identical data;

> window-level and animal-level validation reported side by side on the same dataset, quantifying the inflation that window-level partitioning produces; and the computational and communication costs (feature dimensionality, sampling rate, latency, payload) that determine whether a model can run on a collar. For practitioners the practical results are that a 24-dimensional learned representation is sufficient, that 10–15 Hz sampling retains most of the accuracy, and that among the leading models the choice should be made on update cost and inference pipeline rather than on accuracy differences that do not reach significance." (**Section 4.5, pages 23–24, lines 881–891**)

Comments 7: The manuscript is well written, but some sentences are overly complicated, and you could use more direct language. I added specific comments to the PDF, but I also suggest reviewing the entire manuscript to improve clarity.

Response 7: Thank you. The manuscript has been edited throughout for clarity; the wording points raised in the annotated PDF are addressed individually below.

3.2 Comments in the annotated manuscript

Comments 8: (p. 1, Abstract; highlighted: "rarely isolate individual design choices") Not sure what that means

Response 8: Thank you. We agree that the phrase was unclear and have rephrased it:

> "Comparisons of behavior classification pipelines rarely vary one design factor at a time or report computational cost." (**Abstract, page 1, lines 30–31**)

Comments 9: (p. 1, Abstract; highlighted: "was far more costly,") Is costly the correct term to refer to it? What is the cost?

Response 9: Thank you. We agree that "costly" was imprecise and have replaced it with the quantity itself:

> "Replacing the supervised training objective with unsupervised reconstruction reduced Macro F1 by 0.094–0.127." (**Abstract, page 1, lines 39–40**)

Comments 10: (p. 1, Title; highlighted: "Deployment") I believe computational cost would be a more appropriate term for what was measured.

Response 10: Agree. The title now ends "Accuracy, Validation Design, and Computational Cost" (see Response 5).

Comments 11: (p. 2, Introduction; highlighted: "that intermittent human observation cannot capture") This sentence is not correct, and I suggest rephrased it. Intermittent human observation have captured those disruptions for hundreds of years. Continuous monitoring theoretically enable to detect some issues earlier and more precisely than human observations.

Response 11: Thank you. We agree and have rephrased the sentence along the lines suggested:

> "The time budgets of key behaviors provide information on animal health and welfare that intermittent human observation captures only coarsely; continuous sensing can detect changes earlier and at finer temporal resolution [1]." (**Introduction, page 2, lines 52–54**)

Comments 12: (p. 2, Introduction; highlighted: "In ruminants, rumination and eating time are among the most informative behavioral indicators.") I suggest bring the context here because as it is this sentence is not true. There is many recent studies showing that this behaviour are not relevant for all ruminants. For example, for health issues or productive monitoring in beef cattle. I suggest you revisit recent literature in applied behavioural monitoring of beef bulls and bring the productive context to this sentence.

Response 12: Thank you. We agree that the relevance of these behaviors depends on the production context. The sentence now states the context in which each use is supported and cites a study of beef bulls (Camargo et al., 2024):

> "In ruminants, rumination and eating time are among the most informative behavioral indicators for health surveillance in dairy systems and for health and reproductive monitoring in beef cattle [2,3]." (**Introduction, page 2, lines 54–56**)

Comments 13: (p. 2, Introduction; highlighted: "such estimates may reflect within-animal fit as much as generalization to unseen animals.") I suggest rephrase this sentence. When instances from the the same animal are in training and test sets you cannot infer generalizability in animal behaviour classification. As the major source of variability is between-animal variability and not within animal-variability, if you does not leave animals out of training you are not testing, you are only validating your model.

Response 13: Thank you. We agree and have rephrased the sentence:

> "Because adjacent windows are highly correlated and differences between animals are the main source of variability in behavior classification, such estimates cannot be used to infer generalization to unseen animals [15,16]." (**Introduction, pages 2–3, lines 90–93**)

Comments 14: (p. 2, Introduction; highlighted: "leave-one-subject-out cross-validation, directly tests the latter") LOAO is a validation tool, not a testing tool. It does not test generalization

because you don't have a final model (trained and stable) tested in animals new animals. Please rephrase this sentence.

Response 14: Thank you. We agree and have adopted this terminology: the Introduction no longer describes LOSO as a test, Section 2.7 states what LOSO does and does not evaluate, and "test set" and "test partition" have been replaced by "validation partition" or "held-out partition" throughout:

> "Subject-level validation, such as leave-one-subject-out cross-validation, estimates this generalization directly." (**Introduction, page 3, lines 93–94**)
>
> "…has no subject-level leakage and estimates how the training procedure generalizes to unseen individuals; it does not evaluate a single fixed model on new animals, and its estimates are comparisons among models under matched conditions, not field accuracy." (**Section 2.7, page 11, lines 407–410**)

Comments 15: (p. 2, Introduction; highlighted: "rarely compared on the same dataset") Rahman et al (2018) performed this exactly comparison that informed several studies in the field. I suggest you review this and incorporate this paper in your intro and discussion. 10.1016/j.inpa.2017.10.001

Response 15: Thank you for this reference. Rahman et al. (2018) is now introduced in the Introduction, and the statement has been corrected:

> "Rahman et al. [16] compared within-animal and across-animal partitioning on the same cattle dataset and reported a marked reduction in accuracy under the latter. Since then, however, subject-level validation has remained less common, and few studies have reported both schemes on the same dataset [17,18]." (**Introduction, page 3, lines 94–97**)

The study is also compared with our results in Section 4.3 (see Response 50).

Comments 16: (p. 3, Section 2.1; highlighted: "such as quaternions or rotation matrices. The only signal transformations applied were the per-channel z-score normalization of Section 2.6 and, for the sampling-rate experiments, the resampling of Section 2.8.") I suggest remove this entire excerpt. This is part of your preprocessing and stating that the raw data served as the model input is enough.

Response 16: Agree. The excerpt has been removed:

> "The raw six-channel time series as delivered by the logger served as the model input in every downstream experiment." (**Section 2.1, page 3, lines 128–129**)

Comments 17: (p. 3, Figure 1) I suggest add in the figure the direction of axis in relation to the animals body.

Response 17: Thank you. We agree. Figure 1 now shows the orientation of the sensor axes relative to the animal's body, and Section 2.1 defines them. In preparing the figure we found that the axis assignment in the original Section 2.1 was incorrect: gravity is carried by the X axis, which runs along the collar strap, not by the Y axis. The sentence has been corrected; no result depends on it, because all models received the raw axis-native signals.

> "Figure 1. Wearing position of the collar-mounted six-axis IMU on a Korean black goat and orientation of the sensor axes relative to the body: (a) left lateral view; (b) cross-section of the neck at the collar, cranial view. Arrows indicate the positive axis directions; ⊙ indicates an axis pointing toward the viewer." (**Figure 1, page 4, lines 132–135**)
>
> "The X axis runs along the collar strap (positive ventrally), the Y axis is normal to the neck surface (positive outward), and the Z axis runs along the neck (positive cranially)." (**Section 2.1, page 4, lines 136–137**)

Comments 18: (p. 4, Section 2.1; highlighted: "small subject-specific orientation offsets.") Was that measured to state that this is small? If not I suggest remove the word small. Also, the rotation is usually temporary and not consistent across individuals. Were there individuals with the collar rotated over the entire experimental period or they fixed position after a while? If so this should be noted, otherwise you should recognize that this rotation was temporary.

Response 18: Thank you. The offsets were not measured, so "small" has been removed and the sentence now states that they were not quantified:

> "Collar rotation between individuals and micro-drift during movement introduce subject-specific orientation offsets, which were not quantified." (**Section 2.1, page 4, lines 138–139**)

Collar position was not recorded over time, so we cannot say whether rotation was temporary or persistent in individual animals, and the revised sentence does not assert either. Its possible contribution to between-animal variation is discussed in Section 4.3, with a reference on orientation-independent features for collar-mounted sensors in goats:

> "The models received the native IMU axes without per-subject orientation calibration, so differences in collar fit, resting orientation, and movement-induced rotation can generate animal-specific offsets [19]." (**Section 4.3, page 21, lines 774–776**)

Comments 19: (p. 4, Section 2.2; highlighted: "cameras,") Report the camera model for reproducibility.

Response 19: Agree. The camera model and recording settings have been added:

> "Reference video was captured by ceiling-mounted internet protocol cameras (Roboview 24-3, Global I-net, Gwangmyeong, Republic of Korea; 1920 × 1080, approximately 30 fps), one camera per pen, covering the full floor area with no blind zones." (**Section 2.2, page 4, lines 144–146**)

Comments 20: (p. 4, Section 2.2; highlighted: "Video and IMU timelines") If the cameras were connected to Internet their time drift was close to none. However the devices were not, so there was time drift. The fact that no time drift correction was applied to IMU data should be stated.

Response 20: Thank you. We agree. The paragraph now states how often the timelines were aligned, how IMU timestamps are generated, that no drift correction was applied, and the resulting bound on drift:

> "Video and IMU timelines were aligned manually at the start of each 30-minute recording session. IMU timestamps are generated on the logger with a per-sample sequence number and are therefore unaffected by transmission latency. No clock-drift correction was applied; because alignment was repeated every 30 min, accumulated drift remains below 0.2 s even for a 100-ppm oscillator, under 2% of a 10-s window." (**Section 2.2, page 4, lines 146–151**)

Comments 21: (p. 4, Section 2.2; highlighted: "Ground-truth labels for the two focal behaviors, rumination and eating, were established through a semi-automatic annotation procedure. Trained observers first annotated a seed set of rumination and eating episodes from the reference video … Agreement between the automatic labels and the human video assessment was 99.3% overall (Rumination 98.5%, Eating 99.5%, Other 100.0%).") This can be highly problematic and should be adequately stated in the objectives, and discussed as a limitation of the study. The fact that the ground-truth labels were mostly generated by a video classifier provided by the same company that created the IMU could be a relevant interest conflict. In addition you are including a new source of uncertainty in your data (as all models hold some uncertainty degree). I highly suggest at the very least to mention in your objectives and abstract that the ground-truth labels were mostly obtained by a video classifier. I also suggest you also state as a limitation and discuss this in the discussion section.

Response 21: Thank you. We agree that the origin of the labels must be stated clearly. As explained in Response 2, the automatic labeler is not a video classifier: it operates on the IMU signal, only the human observers used the video, and the manufacturer's role is stated in the Conflicts of Interest. Of the final windows, 58.9% come directly from manual video annotation and 41.1% from the automatic labeler. The label origin is now stated in the Abstract and the objectives, and the first limitation in Section 4.5 now gives the share of automatically assigned windows and the resulting circularity:

> "…with labels from a semi-automatic procedure verified against synchronized video." (**Abstract, page 1, lines 32–33**)

> "Ground-truth labels were produced by a semi-automatic procedure combining manual video annotation with automatic labeling, verified on a stratified sample against synchronized video." (**Introduction, page 3, lines 100–102**)
>
> "About 41% of the final windows were assigned by the automatic labeler, whose KNN stage uses the same handcrafted feature family as Model A; this potential circularity is therefore most relevant to Model A, and if anything biases the comparison against the learned representations that matched it." (**Section 4.5, page 23, lines 841–844**)

Comments 22: (p. 4, Section 2.2; highlighted: "This criterion also limits the influence of any residual timeline misalignment, since an offset of a few seconds affects at most the first and last 10 s window of each episode.") On the other hand this is also a limitation of the model, as in real life animal can express bouts of rumination and eating shorter than 180s. In those cases you don't know how your model will perform. Others have done the same in beef bulls, beef calves, and beef cows, so it is not a enormous limitation but should be recognized. I suggest incorporate the relevant literature here (to support your decision) and discuss this as a limitation in the appropriate section.

Response 22: Thank you. We agree. The criterion is now supported by the study of beef bulls by Camargo et al. (2024), in which very short bouts were also removed, and the corresponding limitation in Section 4.5 now also states what happened to the shorter episodes and cites this study:

> "Only episodes lasting at least 180 s were retained as valid events; very short bouts have likewise been removed in beef bulls, in which they were poorly represented by window-level features [3]." (**Section 2.2, page 5, lines 181–184**)
>
> "Shorter episodes were not discarded but were treated as Other, so the residual class contains a small proportion of brief feeding and rumination bouts. The dataset therefore emphasizes sustained, clearly defined episodes and does not characterize performance on short, fragmented, or transitional events, which are likely the most difficult cases in continuous operation [3]." (**Section 4.5, page 23, lines 851–855**)

Comments 23: (p. 4, Section 2.3; highlighted: "Windows containing more than one behavior class were excluded, so each retained window carries a single unambiguous label.") Once again, there is literature to support these decision in sheep and beef bulls so I suggest referencing to it here. It should also be mentioned as a limitation as in real life those events exist.

Response 23: Thank you. In the studies we could verify, in beef bulls (Camargo et al., 2024) and beef calves (Kim et al., 2024), such windows were retained and assigned to the prevailing behavior rather than excluded; we now cite them as the alternative approach:

> "Windows containing more than one behavior class were excluded, so each retained window carries a single unambiguous label; studies in beef bulls [3] and beef calves [11]

> instead assigned such windows to the prevailing behavior." (**Section 2.3, page 5, lines 197–199**)

We kept the exclusion so that every window in the comparison of feature representations carries an unambiguous label. The resulting limitation, that windows spanning a transition are not characterized, is stated in Section 4.5 (paragraph 2; see Response 22).

Comments 24: (p. 4, Section 2.3; highlighted: "10") Why 10s? There is plenty literature yo support this decision, some of which you already cites throughout the paper. I suggest include these references here.

Response 24: Thank you. We agree. The choice is now supported by references:

> "Window length affects classification accuracy [20]; 10 s is the median window length among accelerometer studies of ruminant behavior [6] and has been used in dairy cows [7] and beef calves [11]." (**Section 2.3, page 5, lines 191–194**)

Comments 25: (p. 4, Section 2.3 title; highlighted: "Construction") You refer to some Python packages (pytorch, numpy) when describing training so I supose this was all done in Python. But you make no reference to any other package or even programming language used. I reccomend specify programing language (with version) and all packages used at every step.

Response 25: Agree. Package and function names are now given at the main steps of the Methods, and Section 2.10 gives the language, the versions and the references (see Response 5).

Comments 26: (p. 5, Section 2.3; highlighted: "subsampled to") Include here you subsampling criteria and package/library used. Were these epochs randomly selected? Did you balanced per subject per day? How much of those represent the manual label epochs an dhow much was comp. vision classification? Include relevant information to ensure reproducibility.

Response 26: Thank you. Windows were selected at random without replacement using NumPy with a fixed seed (Section 2.3, paragraph 2; quoted in Response 3). Balancing was applied at the class level only, not per subject or per day; the resulting numbers of windows per subject and class, with the recording days per subject, are given in the new Table S4. The shares of manual and automatic labels are now reported:

> "Of the 132,887 windows in the final dataset, 78,320 (58.9%), including Other windows, originated directly from manual annotation and 54,567 (41.1%) from the automatic labeler." (**Section 2.2, page 5, lines 170–172**)

Comments 27: (p. 5, Table 1; highlighted: "Final dataset composition") Final dataset composition correspond to the data after subsampling? Oh how many animals over how many

days? Tables and figures should stand alone from the text. I suggest add more details in the table title. e.g., Final dataset composition of retained per-class counts after balancing subsampling processing of XX animals over xx days. Each sample correspond to 10s time-windows.

Response 27: Agree. The caption has been revised as suggested:

> "Table 1. Final dataset composition: retained per-class counts of 10-s windows after high-confidence filtering and class-balancing subsampling, from 19 Korean black goats over nine recording days." (**Table 1, page 5, lines 215–216**)

Comments 28: (p. 5, Section 2.4.2; highlighted: "Model A") Why use LightGBM in such a high dimentional input of 250 features? LightGBM are much more prone to overfit than XGB for example. I suggest add a couple of sentences justifying your choices of LGB and KNN as well.

Response 28: Thank you. We agree and have added a justification of LightGBM:

> "LightGBM was retained because Model A reproduces the pipeline of Kim et al. [11], in which it was the final classifier. With about 133,000 windows, and with row and column subsampling (bagging_fraction = 0.8, feature_fraction = 0.8) acting as regularization, over-fitting of leaf-wise tree growth is limited." (**Section 2.5, page 10, lines 363–366**)

KNN is used in Model C as a non-parametric probe of the learned representation, as already stated where Model C is introduced:

> "Model C tests whether the representation retains discriminative structure under a non-parametric classifier that applies no additional feature weighting." (**Section 2.4.3, page 8, lines 301–302**)

Comments 29: (p. 6, Section 2.4.3; highlighted: "Models B and C") Define Models B and C here. e.g., used in models B (LightGBM downstream classifier) and C (KNN downstream classifier).

Response 29: Agree. The sentence now reads as suggested:

> "Models B (LightGBM downstream classifier) and C (KNN downstream classifier) share a hierarchical three-level feature extractor that maps a raw IMU window of size 250 × 6 to a 24-dimensional learned representation (Figure 3)." (**Section 2.4.3, page 7, lines 251–253**)

Comments 30: (p. 8, Section 2.4.6; highlighted: "LightGBM classifiers are not included in these estimates") Why are they not included in these estimates. If computational cost is an important measure for your study should you have comparable metrics among different classifiers?

Response 30: Thank you. Model size in Table 2 is estimated from neural-network parameter counts, and a tree ensemble has no comparable parameter count. We agree that the classifier stage matters for computational cost: the latency measurements include it, with its share now

reported, and Section 4.4 now discusses its likely memory footprint relative to the extractor, which was not measured:

> "Model B latency covers the deep-learning feature extractor forward pass and LightGBM prediction, the latter contributing 0.12–0.18 ms." (**Section 2.9, pages 12–13, lines 461–463**)
>
> "The tree-ensemble stage of Models B and E, by contrast, comprises up to 1,500 trees of 63 leaves and would likely occupy more memory than the extractor itself, so the end-to-end design of Model D is the more natural candidate for on-collar inference, whereas the modular designs suit gateway- or server-side inference where the classifier can be refitted cheaply." (**Section 4.4, page 22, lines 820–824**)

Comments 31: (p. 8, Table 2 notes; highlighted: "Params refers to neural network parameters; model size is estimated as Params × 4 bytes … For Model D, which classifies internally, the Features column reports the dimensionality of the pooled summary vector rather than of an externally exposed feature vector") Organize this as a proper footnote linking the explanations of acronyms and names in teh columns to the proper explanation. Use the journal guidelines to ensure you use adequate format (eg., superscribed numbers or letters, etc.)

Response 31: Agree. The notes have been reorganized as a general note followed by lettered footnotes (a–c) attached to the cells they explain:

> "Params, neural network parameters; model size is estimated as Params × 4 bytes (32-bit floating point) and excludes the LightGBM classifier." (**Table 2, page 9, lines 346–347**)
>
> "[a] Model A uses LightGBM alone, whose parameterization is not comparable to neural network parameter counts." (**Table 2, page 9, lines 348–349**)
>
> "[b] Dimensionality of the pooled summary vector; Model D classifies internally and exposes no feature vector." (**Table 2, page 9, lines 350–351**)
>
> "[c] Includes the decoder used during reconstruction training; the decoder is discarded at inference, so the deployed encoder matches Model E at approximately 40K parameters." (**Table 2, page 9, lines 352–353**)

Comments 32: (p. 9, Section 2.5; highlighted: "LightGBM classifier") Was any hyperparameter tunning performed in LGBM or KNN? If not, why? Were this hyperparameters came from?

Response 32: Thank you for these questions. No tuning was performed for either classifier. The values were not taken from Kim et al. (2024), who tuned their classifier for calf data with Optuna. Instead, a single configuration was fixed a priori and applied unchanged to Models A, B, E and F (Table 3: learning rate 0.05, 63 leaves, row and column subsampling of 0.8, at most 500 boosting rounds with early stopping), and the KNN classifier of Model C used k = 5 and Euclidean distance (Section 2.5, last paragraph). The reason for not tuning is now stated:

> "The hyperparameters of both classifiers were fixed a priori (Table 3) and not tuned for any model, because tuning the classifier separately for each feature representation would

> confound the feature-extraction contrast that the design is meant to isolate." (**Section 2.5, page 10, lines 367–369**)

Comments 33: (p. 10, Section 2.7; highlighted: "It is retained here because it remains widely used in this literature and provides a reference point for published figures obtained under comparable partitioning.") K-fold is not the standard method for validation in animal behaviour classification. If you decide to keep it here to make it easier to compare with other papers you should state that but you should also mention at some point that for any predictive performance assessment of models the LOSO should be used.

Response 33: Thank you. We agree. Section 2.7 already gave the reason for retaining 5-fold CV; it now also states that LOSO is the appropriate scheme for assessing predictive performance and that 5-fold estimates cannot be used to infer generalization to unseen animals:

> "As the appropriate scheme for assessing predictive performance on new animals, LOSO is treated as the primary result." (**Section 2.7, page 11, lines 410–411**)
>
> "Such estimates cannot be used to infer generalization to unseen animals; they estimate an upper bound on the representational capacity of each model under within-subject sampling." (**Section 2.7, page 11, lines 416–418**)

Comments 34: (p. 10, Section 2.7; highlighted: "test") Validation partition. You did not test your model, you only validated them.

Response 34: Agree. "Test" has been replaced by "validation partition" or "held-out partition" throughout (see Response 14), for example:

> "Leave-one-subject-out (LOSO) validation holds out all data from a single subject as the validation partition and trains on the remaining subjects, cycling through all subjects." (**Section 2.7, page 11, lines 400–401**)

Comments 35: (p. 10, Section 2.7 title; highlighted: "Validation") A section should be added to report the inference tests you used to compare models.

Response 35: Agree. A new Section 2.11 (Statistical Analysis) has been added:

> "Models were compared on per-subject Macro F1 under LOSO at 25 Hz, averaged over the classes present in each subject's held-out data, with each held-out subject as the experimental unit (n = 15). A linear mixed model was fitted with model (A–F) as a fixed effect and subject as a random intercept, using R version 4.6.1 [31] and the lme4 package [32]. Estimated marginal means and Tukey-adjusted pairwise contrasts with 95% confidence intervals were obtained with the emmeans package [33], and paired Cohen's d_z is reported as the effect size for the contrasts discussed in the text. Fold-level differences under 5-fold CV were not tested, because folds are resamples of a single dataset rather than independent experimental units." (**Section 2.11, page 13, lines 489–497**)

Comments 36: (p. 11, Section 2.8; highlighted: "The top-250 feature") I suggest that the name and feature importance of these 250 features to be added as supplemental materials.

Response 36: We thank the reviewer for this suggestion. We have not added the list: the 250 features are indices into a combinatorially generated set of 903 candidates, and Model A serves to reproduce the pipeline of Kim et al. (2024) as a baseline rather than to interpret individual features. The feature families are described in Section 2.4.2, and the full list with importance values is available from the corresponding author on request.

Comments 37: (p. 11, Section 2.8; highlighted: "reused at all lower rates") Have you tested whether selecting features for the appropriate sampling rates rather reuse them change your conclusions?

Response 37: Thank you. We did not re-select features at each rate. As in the original submission, Model A's values below 25 Hz are treated as a conservative estimate for this reason (Section 3.4, paragraph 2), and Section 3.4 now also states that the low-rate difference between Models A and B favors Model A:

> "The Macro F1 difference between Models A and B widened as sampling rate decreased under 5-fold CV, in favor of Model A, from 0.000 at 25 Hz to 0.009 at 5 Hz, but showed no comparable trend under LOSO." (**Section 3.4, page 16, lines 599–601**)

Re-selection would therefore be expected to widen this difference rather than reverse it. The cost of reduced sampling was also of similar size in Models B, D and E, which do not use the selected indices, so the conclusions about the learned representations and the 10–15 Hz operating point are unaffected:

> "Under LOSO, operating at 15 Hz costs 0.003–0.007 in Macro F1 for a 40% reduction in sampled data volume, and 10 Hz costs 0.009–0.012 for a 60% reduction across Models A, B, D, and E." (**Section 4.4, page 22, lines 827–829**)

Comments 38: (p. 12, Section 2.9; highlighted: "Python/NumPy") Review all applications of packages throughout the paper and make the proper reference to them. This is not the standard in scientific literature.

Response 38: Agree. The software is now cited: PyTorch, scikit-learn and NumPy in Section 2.10, which also gives the versions of Python, PyTorch, LightGBM and scikit-learn (quoted in Response 5), and R, lme4 and emmeans in Section 2.11 (quoted in Response 35). LightGBM and SciPy were already cited at first use (Sections 2.4.2 and 2.8).

Comments 39: (p. 12, Section 2.10; highlighted: "2.10.") This description should be either a the beginning of MM or spread across the MM when you first refer to the packages as steps involving them.

Response 39: Thank you. We followed the second option: package and function names are now given at the main steps where they are used (see Response 5). Section 2.10, retitled "Reproducibility and Software Environment", is kept as a short summary of the computing environment to which the latency measurements refer.

Comments 40: (p. 12, Section 3.1; highlighted: "Table 4 summarizes the classification performance of the six models under both validation schemes.") Move this sentence to line 482

Response 40: Agree. The sentence has been moved to the end of Section 3.1, immediately before Table 4:

> "Table 4 summarizes the classification performance of the six models under both validation schemes." (**Section 3.1, page 14, lines 530–531**)

The same change has been made in the other Results sections listed in Response 43.

Comments 41: (p. 13, Section 3.2; highlighted: "The spread between the highest- and lowest-scoring model widened monotonically with class difficulty:") Rephrase this sentence. What do you mean by class difficulty? This is the first time you used this term which seems to be a subjective rank of the authors.

Response 41: Thank you. We agree. "Class difficulty" is now defined at first use by the per-class F1 attained across models:

> "All six models produced the same ordering of class difficulty, defined here by the per-class F1 attained across models: Eating highest, Rumination intermediate, and Other lowest." (**Section 3.2.1, page 14, lines 539–541**)

Comments 42: (p. 14; highlighted: "→") Replace all "→" by "to" in the text. This is acceptable when referring to steps in a written diagram but it is not acceptable when referring to a interval of values.

Response 42: Agree. Value ranges are now written with "to" or an en dash, for example:

> "For Models A and B the 5 to 10 Hz step accounted for more than half of the total 5 to 25 Hz gain, while 20 to 25 Hz contributed no more than 0.003. Across Models A–E the total 25 to 5 Hz reduction ranged from 0.022 (Model A) to 0.043 (Model C) under 5-fold CV." (**Section 3.4, page 16, lines 595–598**)

Arrows are retained only for pipeline steps (Figure 2 caption, architecture descriptions and equations) and for the contrast labels in Table 7.

Comments 43: (p. 15, Section 3.5; highlighted: "Table 8 and Figure") As mentioned above if you refer to the tables and figures at the beginning of the section it is likely that they will placed here instead of after the following text. I suggest move this to line 580 and make the same change across you results section.

Response 43: Agree. The reference has been moved to the end of Section 3.5, and the same change has been made in Sections 3.1, 3.2.1, 3.2.2 and 3.4 and for Table 9 in Section 3.6:

> "Per-sample latencies at batch size 1 are given in Table 8 and Figure 8." (**Section 3.5, page 17, line 632**)

Comments 44: (p. 16, Section 3.6; highlighted: "1,000 animals") Where these values come from if you enrolled 19 animals in your study? If this is a simulation you should report the values you found with your animals and describe in the MM section the simulation parameters and how you did. I don't see any value in scaling up your results. I suggest you report data per animal per day instead.

Response 44: Thank you. We agree. The scaled-up figure has been removed, and the section now reports CPU time per animal-day, derived from the measured per-window latencies:

> "Under continuous monitoring, each animal generates 8,640 windows per day. At the measured per-window latencies, processing one animal-day at 25 Hz corresponds to 373 s of single-thread CPU time for Model A, 23 s for Model B, and 13 s for Model D." (**Section 3.6, page 18, lines 642–644**)

Comments 45: (p. 16, Section 3.6; highlighted: "Speed–Accuracy Trade-off and Aggregate Reduction") LOSO results should be reported instead of K-fold or at least reported alongside to K-fold. Once more, LOSO are significantly more realistic for this type of data so it should prioritized over K-fold.

Response 45: Agree. Figure 9 and Table 9 have been recomputed on LOSO Macro F1:

> "Figure 9. LOSO Macro F1 against throughput for Models A, B, and D across all sampling rates (batch = 1)." (**Figure 9, page 18, lines 647–648**)
>
> "Δ is the change in LOSO Macro F1 relative to the Model A 25 Hz reference." (**Table 9, page 19, line 658**)

Comments 46: (p. 19, Section 4.3; highlighted: "The two validation schemes produced systematically different estimates for every model, with aggregate Macro F1 falling by 0.044–

0.095 from 5-fold CV to LOSO. The inflation follows from how windows are assigned to folds. Allocation is independent of subject identity and temporal position, so data from a single animal appear in both partitions") Revise this excerpt. It's not clear what "The inflation follows from how windows are assigned" means. "Falling by xx-xx from xx-xx" is also odd. You must be clear that the last sentence only apply for K-fold.

Response 46: Thank you. We agree. The excerpt has been rewritten to state the mechanism and to make clear that it applies to 5-fold CV:

> "The two validation schemes produced systematically different estimates for every model, with aggregate Macro F1 decreasing by 0.044–0.095 from 5-fold CV to LOSO. This inflation under 5-fold CV arises from window-level partitioning: allocation is independent of subject identity and temporal position, so data from a single animal appear in both partitions." (**Section 4.3, page 20, lines 729–733**)

Comments 47: (p. 19, Section 4.3; highlighted: "The annotation criterion compounds the effect: because only bouts of at least 180 s were retained, each contributes at least eighteen consecutive 10-s windows, and shuffling distributes windows from one bout across both partitions") This should be discussed as a potential limitation of the model and compare with the literature that did the same or not.

Response 47: Agree. The criterion is now supported by the literature, and the limitation in Section 4.5 now also states what happened to shorter episodes (Section 2.2, paragraph 4, and Section 4.5, paragraph 2; both quoted in Response 22). Section 4.3 compares the resulting inflation with cattle studies that reported both window-level and animal-level validation (see Response 50).

Comments 48: (p. 19, Section 4.3; highlighted: "falling") Replace falling by decreased xx

Response 48: Agree. Replaced by "decreasing by" (Section 4.3, paragraph 1; see Response 46).

Comments 49: (p. 19, Section 4.3; highlighted: "in cattle instrumented at the same neck position and sampling rate.") Rephrase this sentence. e.g., "…in cattle equipped with collars in similar position and sampling rate as the present study" if that is what you meant.

Response 49: Thank you; that is what we meant. In checking the source, we found that the sensors in Wang et al. were described as attached to the neck, without mention of a collar, and that the values we had cited came from the preprint of that study. The sentence now states the placement and sampling rate explicitly and gives the values of the published article:

> "…Wang et al. [15] reported Macro F1 decreasing from 0.89–0.94 under hold-out validation to 0.64–0.76 under leave-cow-out validation in beef cows fitted with neck-mounted accelerometers sampled at 25 Hz, as in the present study." (**Section 4.3, page 21, lines 742–744**)

Comments 50: (p. 19, Section 4.3; highlighted: "As reported in Section 3.1, variation between folds was near zero, whereas variation between held-out subjects was two orders of magnitude larger.") As expected since LOSO is the less inflated CV technique for this type of data. This was describe by many studies in the past I suggest incorporating the relevant literature in beef and dairy cattle to this discussion.

Response 50: Agree. This discussion now includes the cattle literature, and the magnitude in the highlighted sentence has been corrected to "one to two orders of magnitude":

> "Inflation of window-level estimates relative to animal-level validation has been reported repeatedly in cattle. Rahman et al. [16] found F-scores decreasing from 0.81–0.93 under stratified cross-validation to 0.15–0.80 under leave-one-animal-out validation on the same dataset, Riaboff et al. [7] likewise reported lower accuracy under cow-level than under sequence-level validation, and Wang et al. [15] reported Macro F1 decreasing from 0.89–0.94 under hold-out validation to 0.64–0.76 under leave-cow-out validation in beef cows fitted with neck-mounted accelerometers sampled at 25 Hz, as in the present study." (**Section 4.3, page 21, lines 737–744**)
>
> "Variation between folds was near zero, whereas variation between held-out subjects was one to two orders of magnitude larger." (**Section 4.3, page 21, lines 747–749**)
>
> "Between-individual variation of this magnitude is consistent with earlier reports in goats [18] and cattle [15,16], and its dominance carries two consequences." (**Section 4.3, page 21, lines 779–780**)

Comments 51: (p. 19, Section 4.3; highlighted: "The two quantities are not directly comparable: the first is the variation of an aggregate estimate across resamples, the second the dispersion of individual units.") I don't agree with this statement. In the sense that you are inferring predictive performance of different models (which is one objective of the study) they are completely comparable.

Response 51: Thank you. We accept this point. The sentence has been replaced by the point we intended to make:

> "Fold-to-fold variation contains no between-animal information, because every fold draws windows from every subject; it therefore cannot inform generalization to new individuals." (**Section 4.3, page 21, lines 749–751**)

Comments 52: (p. 20, Section 4.3; highlighted: "so its aggregate advantage rests disproportionately on subjects contributing more test windows.") For this reason you should report the distribution of individual and days of you balanced dataset. If they are also balanced per individuals you could isolate this possibility.

Response 52: Thank you; this comment led to a correction. The new Table S4 reports the number of windows per subject and class, with the recording days per subject: windows are not balanced across subjects, and Other windows come from eight of the 15 LOSO subjects. In compiling it we found that the previous per-subject statistic had assigned the other seven subjects an F1 of zero for the absent class; this has been corrected (see Section 2 above). After the correction the rankings by aggregate and by per-subject Macro F1 still differ, but by margins well within the confidence intervals of the mixed model, so we no longer attribute the difference to a particular mechanism and have replaced the highlighted sentence:

> "…because the Other class came mainly from seven fully annotated animal-days, seven of these subjects contributed no Other windows." (**Section 2.7, page 11, lines 403–405**)
>
> "The ranking by aggregate Macro F1 (Model D first) and by per-subject mean (Model A first) differed, but by margins well within these confidence intervals." (**Section 4.3, page 21, lines 770–772**)

Comments 53: (p. 20, Section 4.3; highlighted: "Each subject was also annotated by a single observer without overlap, so systematic differences in annotation practice are absorbed into the same term.") However, no inter-rater reliability test is reported to check whether raters agree with their assessment over time. This should be discussed as a potential limitation.

Response 53: Agree. As explained in Response 3, inter-rater reliability could not be computed and intra-observer consistency over time was not tested; both are now stated as a limitation:

> "Inter-rater reliability of the manual annotation could not be assessed because each animal was annotated by a single observer, and intra-observer consistency over time was not tested." (**Section 4.5, page 23, lines 847–849**)

Comments 54: (p. 21, Section 4.5; highlighted: "Reported metrics should be read as comparisons among models under matched conditions, not as estimates of field accuracy.") This great and should be stated much earlier in the paper. I suggest add similar statement in the MM section so you avoid readers thinking you are over interpreting your results later on.

Response 54: Thank you. We agree. A similar statement now appears in the Methods:

> "…it does not evaluate a single fixed model on new animals, and its estimates are comparisons among models under matched conditions, not field accuracy." (**Section 2.7, page 11, lines 408–410**)

Comments 55: (p. 21, Section 4.5; highlighted: "Three directions") I recommend you reflect on the application of your studies and how your results advance the knowledge in the field.

Response 55: Agree. A paragraph on the contribution and practical implications of the study has been added immediately before this paragraph (Section 4.5, paragraph 5; quoted in Response 6).

Comments 56: (p. 22, Conclusions; highlighted: "Figures obtained under window-level partitioning, here and in the wider literature, are better read as estimates of representational capacity than of expected field performance.") So what is the point if the ultimate goal of ML models is the have high and generalizable predictive performance?

Response 56: Thank you for this question. Window-level figures are not a basis for model selection; the sentence now states why they are reported:

> "Window-level figures should therefore not be used for model selection or for claims about performance on new animals; they are reported here to quantify the inflation and to allow comparison with published results obtained under the same partitioning." (**Section 5, page 24, lines 912–915**)

Comments 57: (p. 22, Conclusions; highlighted: "Operational properties therefore provide a more practical basis for model selection when accuracy differences are small") Are they realistic though?

Response 57: Thank you for this question. We have made the operational properties concrete and stated their limits. Section 4.4 now reports a measured update cost with the assumption behind it and discusses the path to a microcontroller, and the Conclusions name the properties and qualify the latency evidence:

> "On the same workstation, refitting the LightGBM classifier of Model B on precomputed 24-dimensional features took 4.8 s of CPU time (median of three runs), whereas retraining Model D end-to-end took 368 s (36 epochs to early stopping). The comparison assumes that the frozen extractor remains adequate for the new label set, which was not tested." (**Section 4.4, page 22, lines 799–803**)
>
> "A PyTorch model cannot execute directly on a Cortex-M-class device; the usual path is export to ONNX followed by conversion to an embedded runtime such as TensorFlow Lite Micro or CMSIS-NN, with 8-bit quantization. … Measured latency on the target device, and the corresponding power draw, remain to be established." (**Section 4.4, page 22, lines 815–825**)
>
> "Where accuracy cannot separate models, operational properties such as update cost, inference pipeline, and transmitted payload provide a more practical basis for selection; the latency figures reported here characterize relative computational cost on workstation hardware and await confirmation on target devices." (**Section 5, page 24, lines 917–921**)

We hope that the revised manuscript addresses your concerns, and we thank you again for your time and helpful comments.

Sincerely,

Na-Yeon Kim, on behalf of all authors