Moving towards more holistic validation of machine learning-based approaches in ecology and evolution

Authors: Charlotte Christensen+[1,2,3], André C. Ferreira+[1], Wismer Cherono[2], Maria Maximiadi[4], Brendah Nyaguthii[2,5], Mina Ogino[1,2,3], Daniel Herrera C[1], Damien R. Farine[1,3,5,6]

+these authors contributed equally to this work

1. University of Zurich, Department of Evolutionary Biology and Environmental Studies, Winterthurerstrasse 190, 8057 Zurich, Switzerland
2. Mpala Research Centre, P.O. Box 555, Nanyuki 10400, Kenya
3. Max Planck Institute for Animal Behavior, Am Obsterberg, Radolfzell 78315, Germany
4. University of Patras, Department of Biology, Panepistimioupoli Rio, Patras, GR 26500, Greece
5. National Museums of Kenya, Museum Hill, Nairobi, Kenya.
6. Australian National University, Research School of Biology, 134 Linnaeus Way, Acton ACT 2601, Australia

ORCID

CC: 0000-0001-7697-9903
ACF: 0000-0002-0454-1053
BN: 0000-0002-5716-0684
MO: 0000-0002-3780-2942
DRF: 0000-0003-2208-7613

Corresponding authors:

CC: charlotte.christensen@uzh.ch

ACF: andre.marquescondecoferreira@uzh.ch

DRF: damien.farine@anu.edu.au

Conflict of interest statement

The authors declare no conflicts of interest.

Authors contribution

CC, ACF and DRF conceptualised the study. CC and BN led the data collection of the accelerometer and video data in the field. CC wrote the ethogram and CC, WC and MM annotated the videos. DH corrected the time alignment between annotation and video data. CC trained the random forest model and performed the statistical analyses for the inter-observer agreement and the biological validations. DRF conducted the simulations. ACF collected the data and trained the deep learning models for Box 1. MO conducted the food supplementation experiment for one of the biological validations. CC and ACF wrote the first draft of the manuscript with the input of DRF. All authors read, improved and approved the final version of the manuscript.

Acknowledgements

We would like to thank the Vulturine Guineafowl Research Programme field team: Monicah Wambui, Kennedy Kipkorir, John Wanjala and Mary Waithira Ngugi for their work in deploying tags, downloading biologging data and collecting videos that were annotated for model training. We would like to thank Mpala Research Centre for logistical support and allowing us to work on the conservancy. We would like to thank National Museums of Kenya (NMK), the National Commission for Science, Technology and Innovation (NACOSTI) and the Wildlife Research and Training Institute (WRTI) for approval of our work. CC and ACF were funded by European Research Council (grant agreement no. 850859 awarded to DRF). CC was supported by a UZH post doc grant (no. K-74312-02-01). MO received additional funding from the Swiss Federal Commission for Scholarships. DRF was funded by an Eccellenza Professorship Grant of the Swiss National Science Foundation (Grant Number PCEFP3_187058).

Data availability

Data and code will be made available on Figshare once the manuscript is accepted for publication.

Abstract

1. Machine-learning (ML) is revolutionizing the study of ecology and evolution, but the performance of models (and their evaluation) is dependent on the quality of the training and validation data. Currently, we have standard metrics for evaluating model performance (e.g., precision, recall, F1), but these to some extent overlook the ultimate aim of addressing the specific research question to which the model will be applied. As improving performance metrics has diminishing returns, particularly when data is inherently noisy, biologists are often faced with the conundrum of investing more time in maximising performance metrics at the expense of doing the actual research. This leads to the question: how much noise can we accept in our ML models?
2. Here, we start by describing an under-reported source of noise that can cause performance metrics to underestimate true model performance. Specifically, ambiguity between categories or mistakes in labelling of the validation data produces hard ceilings that limit performance metric scores. This common source of error in biological systems means that many models could be performing better than the metrics suggest.
3. Next, we argue and show that imperfect models (e.g. low F1 scores) can still useable. We first propose a simulation framework to evaluate the robustness of a model for hypothesis testing. Second, we show how to determine the utility of the models by supplementing existing performance metrics with 'biological validations' that involve applying ML models to unlabelled data in different ecological contexts for which we can anticipate the outcome.
4. Together, our simulations and case study show that effects sizes and expected biological patterns can be detected even when performance metrics are relatively low (e.g., F1 between 60-70%). In doing so, we provide a roadmap for validation approaches of ML models that are tailored to research in ecology and evolutionary biology.

Keywords

Deep learning; Hypothesis testing; Machine learning; Model training and evaluation; Performance.

Introduction

Advances in technology have dramatically enhanced our capacity to gather data on animals (Kays *et al.* 2015), giving us insight into where animals go (e.g., GPS: Cagnacci *et al.* 2010), what they are doing (e.g., tri-axial accelerometers: Brown *et al.* 2013) and with whom they interact with (e.g., proximity tags: Ryder *et al.* 2012). This drastic increase in data collection has also been accompanied with advances in computer science, namely in machine learning, allowing us to process and analyse large amounts of complex data. For example, machine learning has been used to classify species of animals from millions of pictures from camera traps in a fraction of the time of a human performing the same task (Willi *et al.* 2019; Norouzzadeh *et al.* 2021; Whytock *et al.* 2021). Nevertheless, hidden beneath the published studies that successfully used machine learning models are potentially many others where low model performance (e.g. low F1 scores) have made them less appealing for publication. This has likely resulted in a publication bias comprising of artificially high model performance and an over-optimism about ML performance (e.g., Jennions & Moeller 2002; Saidi, Dasarathy & Berisha 2024), as well as substantial lost research time when high model performance has seemed unattainable.

Low scores on performance metrics (see: Rainio, Teuho & Klén 2024 for a review on commonly used evaluation metrics) understandably lead researchers to invest in a time-consuming process of refinements. These refinements can include testing different hyper-parameters, model architectures, or machine-learning model approaches. However, the improvements yielded through what is effectively trial-and-error exploration are typically marginal (Probst, Boulesteix & Bischl 2019; Sipper 2022). Increasing the training dataset is usually the most effective way of improving models' performance, which can motivate a large time and/or financial investment. Collecting more data is, however, often not viable either due to the data collection period being limited in space and time (e.g., within a delineated field season: Shuert, Pomeroy & Twiss 2018; Christensen *et al.* 2023) or because certain biological phenomena or species are hard to observe making it challenging to collect more data (Hounslow *et al.* 2019; Zhong *et al.* 2021). Moreover, even if more data are collected, the collation and annotation of training data is time and resource intensive (e.g.,

labelling photos of animals with the species names: Norouzzadeh *et al.* 2021; or labelling hours video second-by-second: Christensen et al., 2023). Improving model performance can therefore be a frustrating and time-consuming endeavour.

We propose that more effective use of researcher time can be achieved by better understanding the sources of model (apparent) errors, how these translate into model performance when deployed, and how model performance affects (or not) hypothesis testing. When comparing different methods and tweaking parameters fails to make the desired improvements, poor model performance is typically ascribed to insufficient or non-representative training data or to high class similarity (e.g., being groomed while sitting, lying or standing: Christensen *et al.* 2023). In some cases, categories that are being misclassified for one another may be merged into broader categories (Ladds *et al.* 2017) without jeopardizing the initial hypotheses (e.g., if the act of being groomed rather than postural data is of importance: Christensen *et al.* 2023). However, the true blame for (apparent) low performance may not lie solely at the feet of incomplete training data or the similarity of different classes, but rather in the assumptions about our training data. While ML models (in ecology) are typically used for classification problems (e.g. categorising distinct behaviours), natural data are often continuous. Behaviours also often entail transitional states (Reyes-Ortiz *et al.* 2016; Riaboff *et al.* 2022), such as switching from feeding to moving, making it hard to delineate exactly when in time the change occurred. Subtle differences between seemingly similar categories (e.g. an animal having different intentions when raising its head for predator vigilance versus looking for the next food patch) can also make labelling ambiguous, even for experienced observers. The inherent ambiguity and noise associated with biological systems naturally introduce inaccuracies in the labelled data used to train and (importantly) evaluate ML models.

Inaccuracies in the labelling process represent an under-appreciated challenge for our ability to evaluate models. While it may seem logical that inaccuracies limit models from achieving maximal performance because the models have to learn from data with errors, it appears that ML models can be resilient to some mislabels (Rolnick *et al.* 2017) (Box 1). Instead, what has been almost completely overlooked is that mislabels during the training process hinder our ability to *evaluate* model performance, rather than limiting the true performance of the model. This is because a model that predicts a correct label will still get penalised (i.e. reported performance in terms of accuracy,

recall, F1 score, etc.) when human-assigned labels in the validation set contain errors (Box 1). Errors in the labelling, which are common in biological data, can therefore make it impossible for models to achieve perfect performance metric scores during evaluation, even if they are in fact performing very well. Thus, accurately determining a model's performance can and should go beyond simply relying on performance metrics.

The fact that models sometimes underperform (particularly in terms of classical metrics), should not necessarily hinder their application in ecology. Contrary to computer scientist that are motivated to push the boundaries of the state of the art in ML by maximizing model performance (Birhane *et al.* 2022), biologists are primarily interested in using ML models as a tool to facilitate data processing and ultimately test biologically relevant hypotheses (Hilborn & Mangel 2013). In ecology and evolution, (statistical) model performance is typically evaluated in terms of its ability to provide robust tests of hypotheses (and estimating corresponding overall effect sizes), with much less emphasis put on the ability of models to make predictions for specific data points (Houlahan *et al.* 2017). Because of the noisy nature of biological data, our hypothesis testing tools, e.g., linear regression models (Bolker *et al.* 2009) or Bayesian inference (Ellison 2004), are already built to handle (random) errors in our data (Young 2018). Our ability to statistically account for noise (in this case errors in prediction) pulls into question whether—in the context of biological hypotheses testing—performance metric maximization (e.g. maximizing F1 scores) is a useful pursuit, or whether there are more informative ways of deciding when we should stop the training process and move towards model deployment for hypothesis testing. Here we argue that while current performance metrics are useful during model training and to compare different approaches, biologists should look beyond pure metric maximization. Instead, we should ask whether a model is able to perform well enough to confront the biological hypotheses of interest with the data that the model predicted.

In this paper, we outline new perspectives and solutions on how to assess ML model performance for the purpose of hypothesis testing. As a case study, we focus on the identification of behaviour from accelerometer (ACC) data collected on a free-ranging population of vulturine guineafowl (*Acryllium vulturinum*). Combining ACC data with ML approaches to infer behaviour is a powerful way of determining the behaviour of animals without the need to observe them directly. We first use a popular ML approach

(random forest models) to classify behaviours using labels assigned from videos. To better understand sources of error and how these affect model performance metrics we compare the behavioural labels from different annotators. Our comparison shows that annotator disagreement is one source of error in the evaluation process that can result in the underestimation of the performance of the model for some behaviours (see also Box 1).

Our models, like most models in ecology, still produce a significant number of true prediction mistakes that could not be directly attributed to mislabels in the validation dataset. We therefore also investigate whether we can still use our model to test our biological hypotheses of interest. First, we develop a simulation framework to evaluate whether imperfect behavioural classification models can still generate useful data for hypothesis testing. In doing so, we shift the focus from model performance metrics to the contribution that the model makes towards detecting true biological effects. We highlight that this is critical for studies in the biological sciences because prediction errors will have different impacts depending on the hypothesis that is being tested (Box 2). Finally, to confirm the utility of ML models, we propose an additional 'biological validation' procedure that uses model predictions on unlabelled data to test simple hypotheses that we know *a priori* to be supported in our study system. In this way, we can evaluate the model's ability to predict data to test biological relevant questions before using it to test the hypothesis of interest (see: Ferreira *et al.* 2020a for a similar approach in social network analyses).

Case study methods

Collecting, labelling, and matching video data with ACC data

We use data from a population of vulturine guineafowl that are part of a long-term project at the Mpala Research Centre, Laikipia, Kenya. Birds are fitted with solar-powered GPS and accelerometer tags (15 g Bird Solar, e-obs Digital Telemetry, Grünwald, Germany, for more details see: Papageorgiou *et al.* 2021) which can be remotely re-programmed to collect accelerometer and GPS data (see Table S1 for details on ACC birds). Videos of n=14 birds were collected (using Panasonic HC-V777) in the field from a stationary car. At the start of the video, the date and exact GPS-time stamp were filmed to time-match the video with the ACC data. Videos were

annotated by two trained observers (co-authors: *name 2 redacted*, *name 3 redacted*) using the ELAN software (version 6.4) following an agreed upon Ethogram (Table S2), consisting of four modes of locomotion (walking, trotting, running, flying), two modes of foraging (pecking at ground, nipping at foliage), two modes of self-directed care (preening and dustbathing), resting (seated, lying or in combination with preening or pecking), and standing alert. Note that vulturine guineafowl are almost entirely terrestrial, flying only rarely. Event behaviours (e.g., social interactions) and transitions between behaviours were categorised as 'other' during the video annotations. We annotated 13.4 hours of video (n=258 videos).

In R (R version 4.1.1; R studio 2024.04.2), after merging accelerometer data to video data, misaligned labels were corrected by one of the co-authors (*name 4 redacted*). Some videos (n=18) did not have distinct enough accelerometer profiles to confidently confirm or correct the alignment between ACC data and annotations, and these were removed from the training data. These removals, together with periods where no bird was present in the video (1.5 hours) resulted in a dataset of 11.9 hours across the 14 birds. Annotated behaviours were merged into biologically similar categories or into a category called 'other' if behaviours were rare (drinking, social interactions) or not of interest (transitions) (Figure S1). This resulted in 7 behaviours for training the random forest model (Table S3 for sample sizes). After matching annotated video data to 1Hz ACC data (see below), seconds that included more than one annotation were assigned the category that took up the larger proportion of that second.

Metric calculation and random forest model fitting

We summarised each second of 20 Hz ACC data into 1 Hz data with 14 variables: static acceleration (Xst, Yst, Zst), dynamic acceleration (dyX, dyY, dyZ), raw VeDBA, smoothed VeDBA using a running mean of 2 seconds, and the first and second power spectrum densities for all three axes using Fast-Fourier transformation also at an interval of 2 seconds (adapted from: Fehlmann *et al.* 2017; Christensen *et al.* 2023). The random forest models were built the R package 'randomForest' using the default setting of n=500 trees and mtry=3.

We ran a random forest model with seven behaviours (Table S3). We trained the random forest model using a leave-one-individual-out (LOIO) approach (Chakravarty *et al.* 2019). This approach avoids overestimating model's performance, which

originates from autocorrelation between the training and validation data and gives a more accurate reflection of the performance of the model when generalised to data from birds that were not included in the original training dataset (Ferdinandy *et al.* 2020). During the validation procedure, for the models in which two behaviours took place within the same second, the prediction would be labelled as 'correct' as long as it corresponded to one of those two behaviours. A confusion matrix was produced (Table S4) and precision, recall and F1 scores were calculated using the standard formulas (see Supplementary material). Finally, a random forest model containing all of the labelled data was trained to be applied to classify behaviours from new data (see '*Biological validation can directly test the utility of imperfect models*').

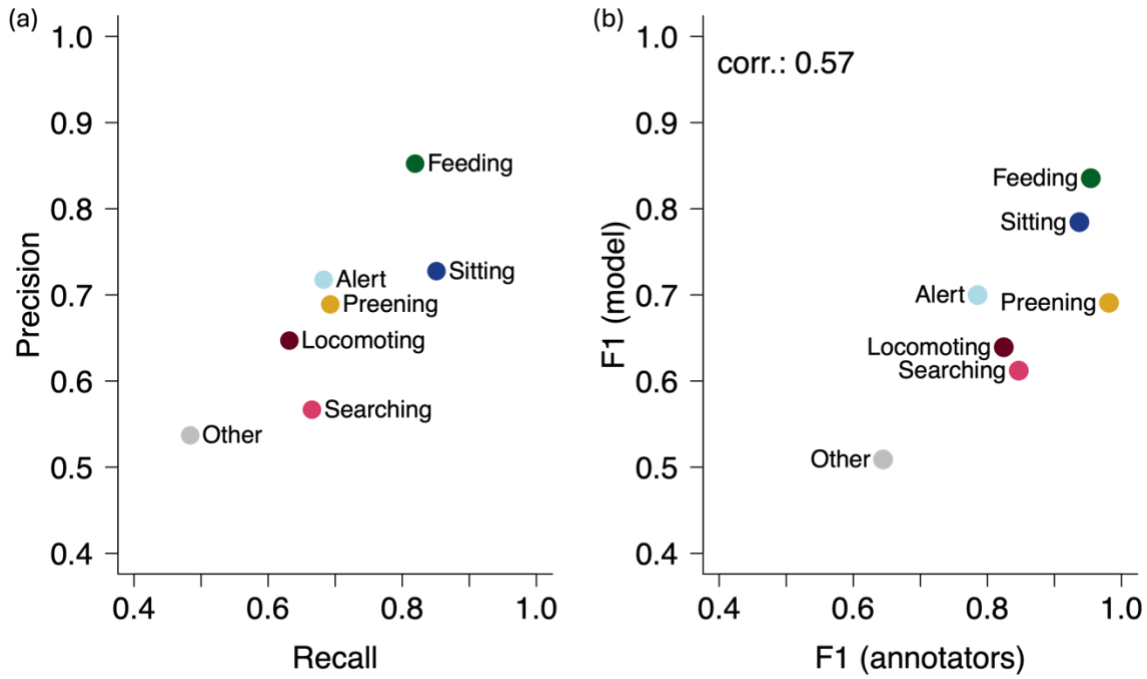


***Figure 1*: Evaluation of the performance of random forest model using standard evaluation metrics is capped by mislabels during the training process** (a) Precision and recall for 7 behaviours. (b) Correlation between the F1-scores of the annotators (i.e., inter-observer agreement) and the model F1-scores (note axes start at 0.4, rather than 0.0 for easier visualisation)

Inter-observer agreement illustrates imperfect labelling

The reliance of human labelling— the current 'gold standard' [20]—has largely overlooked the issue of ambiguity and errors when creating the 'ground truthed' data, at least in ecology (but see examples from computer science: Sheng, Provost & Ipeirotis 2008; Yan *et al.* 2014). Here we demonstrate how quantifying inter-observer agreement can illustrate why there is a 'cap' on our model performance metrics, leading us to potentially underestimate our models. To estimate inter-observer agreement, we used data from 62 minutes of video encompassing all 7 behaviours that were labelled by the person who wrote the ethogram (*name 1 redacted*) and one of the two trained observers (*name 2 redacted* and *name 3 redacted*). We built a confusion matrix (Supplementary Material) and calculated F1 scores with the labels from *name 1 redacted* considered as the "ground truth". As expected, we did not find a perfect agreement between annotators (Table 1) even though they followed the same ethogram (Table S2) and were trained by the same person. On average annotators disagreed on 14.7% of the labels. We might predict that behaviours with lower inter-observer agreement (as a proxy for mislabels) are more difficult to classify due to errors in the training and validation datasets. Indeed, we find a moderate positive correlation (0.57) between F1 score and the agreement between observers (Fig 1; Table 1). However, inter-observer disagreement (mislabels) does not completely explain some low F1 scores, particularly for behaviours with relatively high concordance (e.g. 'Preening' has the highest concordance but only ranks 4th on the model's F1 scores; Table 1). Such low performance can be a result of low sample size or because some behaviours that are easily distinguished on video can generate very similar streams of accelerometer data (Shuert, Pomeroy & Twiss 2018; Chambers *et al.* 2021). Further, we find that both annotator disagreement and model error are class dependent (some classes are never confused, and almost all confusions occur among a limited set of classes; Table S4 and Table S5).

Table 1. Labelling errors correlate moderately with the performance of ML models. Precision (model ability to minimize false positives), Recall (model ability to minimize false negatives) and F1 score (the harmonic mean of Precision and Recall) of the random forest model, and F1 score for annotators (as a measure of inter-observer agreement). Macro F1 is the average F1 score across all 7 categories.

	Precision	Recall	F1 (model)	F1 (annotators)
Alert	68.3	71.8	70.0	78.5
Feeding	81.9	85.2	83.5	95.4
Locomoting	63.2	65.0	63.9	82.4
Other	48.4	53.7	50.9	64.4
Preening	69.3	68.9	69.1	98.1
Searching	66.5	56.7	61.2	84.6
Sitting	85.1	72.8	78.4	93.7
Macro F1			**68.1**	**85.3**

Imperfect models can still be powerful for hypothesis testing

Our end goal when using machine learning is often not the predictions themselves, but rather to use these predictions for hypothesis testing. In this section, we develop a simulation framework that allows us to test the utility of machine learning models for hypothesis testing—specifically determining the ability to detect a given effect size in the presence of false positives or false negatives. We start by generating an observed second-by-second behaviour dataset that spans 10 days (collecting 7200 samples per day, e.g. 2 hours of 1 Hz accelerometer data). For each simulated second, we allocate a behaviour using the real probabilities of these behaviours taking place (based on the prevalence in the labelled guineafowl dataset), and then generate inferred behaviours by sampling from the confusion matrix (i.e. the probability that the real behaviour is inferred as each of the other behaviours) from the random forest model (Table S4). For the purpose of this simulation, we considered that the confusion matrix is not influenced by annotator disagreement in the labels (i.e. it represents “ground-truth” confusion matrix). To create a biological effect, we alter the frequency at which a specific behaviour was expressed during five of the days (e.g. increasing ‘searching’ from one time period to another by between 0 to 20% of the time budget in increments of 1.25% while reducing all other behaviours proportionately). We repeat this process 100 times for each effect size. For each study (a simulated ‘observed’ dataset), we regress the rate of the focal behaviour against treatment (period 1 vs period 2) and

report the true effect (i.e. the effect size from the simulated data) and the observed effect (i.e. the effect size after classification error is introduced using the confusion matrix), as well as the P value for each model. Because misclassifications are non-randomly distributed, we expect that one possible source of false positives would be in a behaviour that is commonly mistaken with the observed behaviour (e.g. in the case of 'searching' the most commonly confused behaviour is 'locomoting'; Fig. 2e). Thus, for each study we also calculated the effect on the behaviour that is most commonly mistaken with the behaviour of interest (Fig. 2; Table 2 for 'searching'; Table S6 for remaining behaviours). We predict that misclassifications would lead the to an underestimation of the focal behaviour and that this could lead to an overestimation of the most commonly confused behaviour(s).

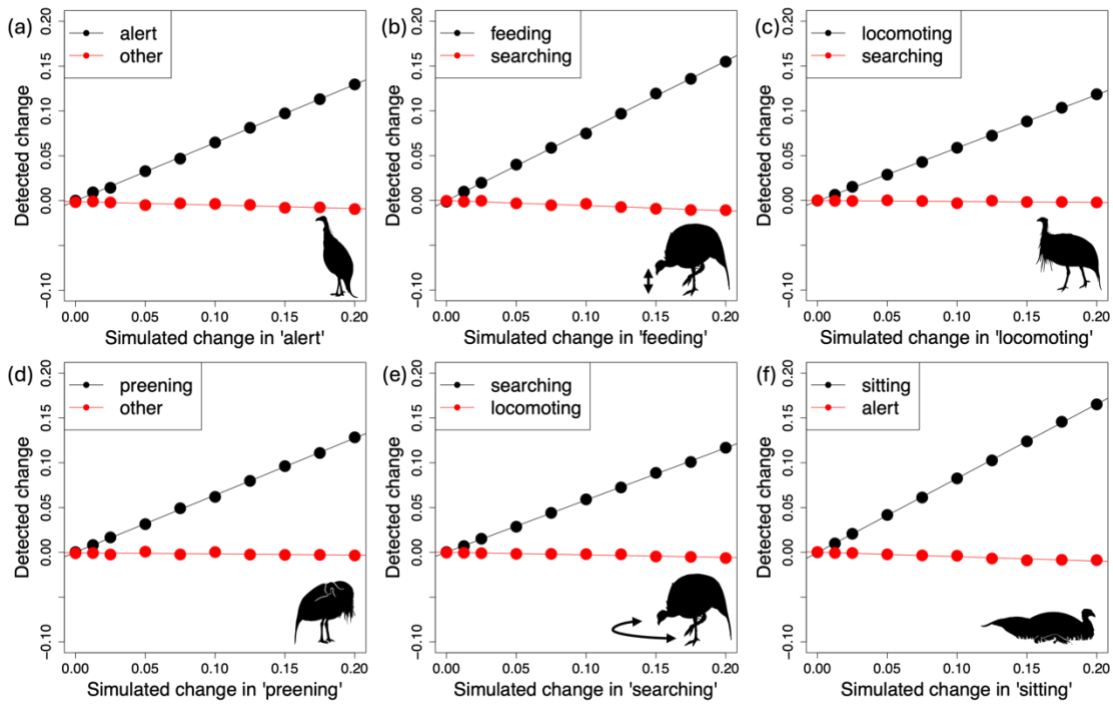


***Figure 2*. Behaviours that are increased in a dataset ('simulated change') can be detected despite classification error ('detected change').** As the simulated frequency of a behaviour is increased (0.0 to 0.2 in increments of 0.0125; x-axis) so does the inferred rate of that behaviour from the model (y-axis, black lines and points). We find that while the frequency is underestimated, we detect no positive increase in the most frequently confused behaviour (red lines and points). Silhouettes by James Klarevas-Irby.

Table 2. Imperfect ML models can be used to make robust biological inferences. Results from simulation where the propensity of 'searching' is increased between two periods in the data by a given amount (simulated change, see Fig. 2e). Detected mean change captures the mean of the change between the two periods over all 100 simulated studies. Proportion of P < 0.05 is the proportion of the simulated studies that detected a statistically significant change. Results for the other behaviours are presented in the Supplementary Material (Table S6).

Simulated change	**Detected mean change**	**Proportion of *P* < 0.05**	**Detected mean change**	**Proportion of *P* < 0.05**
	Searching (true behaviour)		*Locomoting (confused behaviour)*	
0	<0.001	0.1	<0.001	0
0.0125	0.008	0.7	<0.001	0
0.025	0.015	1.0	<0.001	0
0.05	0.027	1.0	-0.001	0
0.075	0.045	1.0	-0.003	0.1
0.01	0.06	1.0	-0.002	0.0
0.125	0.073	1.0	-0.004	0.1
0.15	0.088	1.0	-0.003	0.4
0.175	0.104	1.0	-0.005	0.4
0.2	0.118	1.0	-0.007	0.9

Our simulations show that although our models produce some noisy estimates, they can still be robustly used for testing biological hypotheses (Table 2; Table S6; Fig. 2). We found that even small changes in behaviour can still be reliably detected (i.e. high rates of significant models). Importantly, the behaviour most commonly confused with 'searching' (i.e., 'locomoting') does not increase meaningfully as 'searching' increases (effect sizes remain 1-2 orders of magnitude smaller than those of the real effect),

suggesting that biased misclassifications (in our example) should not be expected to be a high source of false positives. Misclassified behaviours instead show a slight decrease, corresponding with the reduced time spent to make way for an increase in the focal behaviour. Further, models with no effect (simulated change = 0) also have low rates of false positives (rates of significant models is <0.1). Finally, it is worth noting that classification errors can result in underestimates of the true effect size for behaviours with low F1 scores (e.g., searching F1 = 61.2; Table 1). From our simulations, we conclude that while imperfect models can be useful, reporting the results of models should consider the relative strengths of the effects detected. This is a common issue in the era of big data, as very small effects can be highly significant simply because of large datasets, which is one reason for P values being unreliable (Greenland *et al.* 2016; Colquhoun 2017).

Biological validation can directly test the utility of imperfect models

One major challenge for applications of machine learning in ecology and evolution is the general necessity to apply models to data collected outside of the scope of the training data. For example, we may fit tags to new individuals during the study, use data collected after the training period has ended, or test hypotheses on data collected in different locations to those where we collected the training data. Few studies on wild animals have quantified the validity of using data collected outside of the scope of the training data (but see: Wang *et al.* 2015; Brewster *et al.* 2018; Giese *et al.* 2021; Christensen *et al.* 2023). Here we propose an approach that can both (i) test if a model allows for the identification of biologically sound patterns in true field conditions (i.e. extending the validation beyond simulation presented above), and (ii) test if a model is likely to remain useful once applied to unseen data. Our approach is based on biological validation, i.e., confirming that the data can identify effects known *a priori* to exist. For example, walking long distances is (in most animals) incompatible with long periods of resting or foraging. We should also expect higher rates of foraging behaviours in areas where animals are known to forage compared to roosting areas. Here we illustrate the concept of biological validation for ML models by developing four cases in which we test a “simple” hypothesis as a means of examining the performance of a model at addressing a real biological problem. While these “simple”

hypotheses and *a priori* knowledge will always be species specific, some or all of the examples proposed here could easily be adapted to most study systems. We also note that not all behaviours may have a suitable "simple test" and that this may lead to issues of 'coverage' (Pei *et al.* 2017), which means not all branches of a model may be biologically validated (e.g., we have no clear prediction for the category 'Other' in our case study). However, we cover the behaviours of interest (i.e., that will be used to test hypotheses in future studies).

Biological validation 1: vulturine guineafowl spend more time foraging in proximity to food supplements

To test the detection of feeding versus searching behaviours (Table S4), we take advantage of food supplementation that was available to the population between 3rd June and 12th July 2023. Small patches of distributed millet grain (5 by 5 meters) were visited by vulturine guineafowl early in the morning. We expect that feeding behaviour should be detected more frequently in close proximity to a patch. By contrast, the easily-confounded searching behaviour (associated with slow-paced walking and scanning the ground for food) should not. For n=8 birds, we extracted the GPS data from 06:00 to 07:30 and calculated the proximity to a foraging patch (< 15 m or > 15 m). We then use generalized linear mixed effect models (R package 'lmerTest') with 'proportion behaviour per minute' fit as a binomial response variable and proximity (< 15 or > 15 m) as a categorical fixed effect, and 'bird ID' and 'Date' as random effects. Of the total 429 hours of GPS data that satisfied the time criteria, 38 hours were within 15 m of the foraging patch with the remaining (391 hours) being outside the 15 m perimeter. The results (Table 3; Fig. 3a) confirm that guineafowl in closer proximity to the millet patch spend more time foraging. By contrast, there are very small (and negative) changes in searching behaviour.

Table 3. Biological validation demonstrating that the random forest model can detect an increase in foraging close to a known food patch. Results of a generalized linear mixed model on the inferred proportion of time spent (a) feeding and (b) searching when in proximity (<15 m) of a known food patch relative to the baseline (>15 m). Data for 8 birds across 41 unique dates with 25747 1-minute blocks.

Predictor	Estimate	SE	z-value	*p*
(a) Feeding				
Patch (<15 m)*	1.516	0.007	215.53	<0.001
(b) Searching				
Patch (<15 m)*	-0.128	0.009	-14.74	<0.001

Biological validation 2: vulturine guineafowl spend more time locomoting when making larger displacements

To test the ability to distinguish between locomoting and searching (Table S4), we use the GPS data to calculate the distance travelled over 10-minute periods during which the accelerometer was also collecting data. We then match this to the time spent locomoting and searching in those same 10 minutes. We expect that moving larger distances should be strongly and positively correlated with locomoting (travelling in a directed fashion). By contrast, we do not expect to find support for such a relationship for searching behaviour (travelling while looking for food). We fit generalized linear mixed effect models with proportion of the 10-minute window spent doing the behaviour of interest ('locomoting' or 'searching') as the response variable and GPS-distance travelled was the predictor variable (scaled from 0-1 using R package 'scales'), controlling for bird ID and Date as random effects (in a binomial model). We use the whole dataset from the same periods as biological validation 1. Our analysis confirms that larger GPS displacements positively predict the proportion of 'locomoting' detected, but not searching (Table 4; Fig. 3b).

Table 4. Biological validation demonstrating that the random forest model can detect greater locomotion during periods when individuals make large displacements. Results of generalized linear mixed models on the inferred proportion of time spent (a) locomoting and (b) searching as a function of the distance travelled for 8 birds across 41 unique dates with 2830 10-minute blocks.

Predictor	Estimate	SE	z-value	*p*

(a) Locomoting				
Distance (scaled)	3.536	0.013	267.61	<0.001
(b) Searching				
Distance (scaled)	-0.036	0.016	-22.38	<0.001

Biological validation 3: vulturine guineafowl spend more time 'sitting' at night than during the day

To test the ability to distinguish between 'sitting' and 'alert' (Table S4), we test whether night time is associated with an increase in 'sitting' relative to a matched day-time window. Vulturine guineafowl spend the night at communal roost sites (Papageorgiou *et al.* 2024). While some wakefulness is expected at the roost, the night should be generally associated with more 'sitting' (stationary) relative to morning hours when vulturine guineafowl are observed to forage with intermittent bouts of alertness. The tags of three vulturine guineafowl were reprogrammed between 25th May 2023 and 6th July 2023 to collect data from 18:00 to 11:00 the next day. We use these data to test whether the proportion of time spent sitting was higher between 00:00 and 02:00 (middle of the night) than 07:00 and 09:00 (morning activity hours when birds are observed to forage on the open glades), controlling for 'bird ID' and 'Date'. We show that birds spend more time 'sitting' at night than during the morning hours (Table 5; Fig. 3c).

Table 5. Biological validation demonstrating that the random forest model can detect changes in sitting and alertness as a function of the time of the day. Results of generalized linear mixed models on the inferred proportion of time spent (a) sitting and (b) alert at night (00:00 to 02:00) versus during the morning (07:00 to 09:00; baseline) for 3 birds across 12 unique dates with 5844 1-minute blocks.

Predictor	**Estimate**	**SE**	**z-value**	***p***
(a) Sitting				
Time of day (Night)	5.272	0.034	154.857	*<0.001*

(b) Alert				
Time of day (Night)	-0.203	0.109	-18.507	*<0.001*

Biological validation 4: vulturine guineafowl show higher within-group than between-group synchrony in their activity

In some cases, *a priori* predictions about which behaviours will increase or decrease could be difficult to make, or experimental manipulations may not be feasible. In the case of group-living animals (such as the vulturine guineafowl), we can use the expectation that individuals in closer proximity (e.g. those from the same group) should be temporally more synchronised in their behaviours than dyads that are farther apart. These patterns are likely because individuals in close proximity experience more similar ecological forces, while maintaining cohesion also requires some synchronisation of activity (Davis, Crofoot & Farine 2022). with the caveat that some behaviours might also happen asynchronously, (e.g., alternating bouts of sentinel behaviour in meerkats: Santema & Clutton-Brock 2013)). We use this prediction to test the performance of our model by taking the behavioural data from one vulturine guineafowl and time matching to the data from a group member and a non-group member for 8 consecutive days (between 3rd and 10th of June 2022). For every second in these data, we calculated whether the behaviour was the same or not. Note that all three individuals live in close proximity, have overlapping home ranges, and experience largely identical ecological conditions. The model outputs confirm a substantially greater increase in synchronisation among individuals in the same group (mean±SD 28.2±0.04% of the time) than among individuals in different groups (0.05±0.01% of the time, Figure 3d).

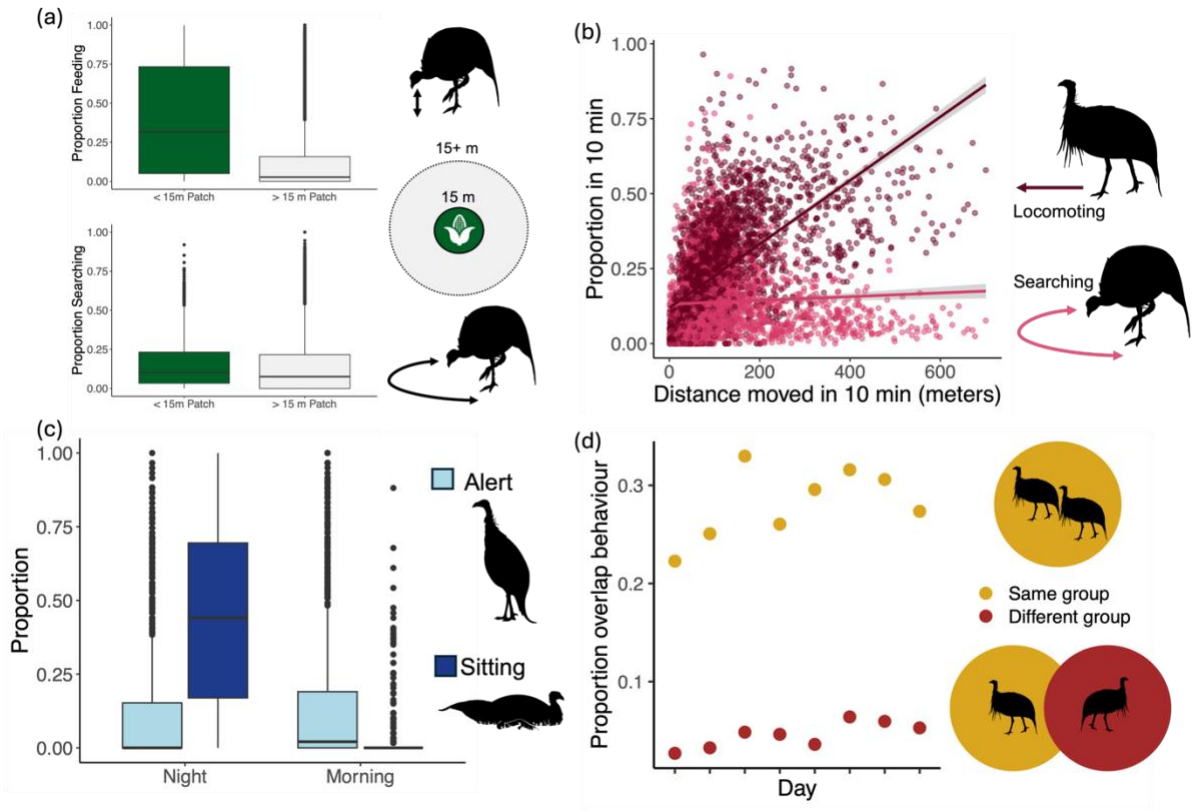


Figure 3. Biological validation provides confirmation that the machine learning model is capable of testing biologically relevant hypotheses. (a) A higher rate of feeding but not searching is detected around a supplemented millet patch. (b) A higher GPS displacement corresponds to a higher proportion locomoting but not searching. (c) A higher rate of sitting (resting behaviour) is detected at night than during the day, whereas a higher rate of alert behaviour (also immobile) is detected during the day. (d) An individual has 5 to 8 times higher overlap in second-by-second behaviour with another individual moving within the same group compared to an individual in a different group. Silhouettes by James Klarevas-Irby.

Discussion

In this work we highlight the difficulty in accurately estimating performance metrics of ML models based on biological data, and how this can cause misguided investments or undue critiques of models. We first demonstrate a moderate correlation between the ability of humans to classify data and the performance of the model at making the same classification, suggesting that errors in the validation dataset create a ceiling on metric performance. In addition, we suggest that while models are likely to always

output a proportion of erroneous predictions, this may not necessarily hinder the utility of the model. As biologists, our aim is often to test hypotheses rather than make predictions *per se*, and our simulations suggest that models can still be useful even if imperfect. Finally, we propose methods for implementing validation approaches that align with the ultimate aim of using models for hypothesis testing, demonstrating that our own 'imperfect' models can reliably detect a number of predicted patterns in behaviour. In doing so, we provide tools for developing and evaluating models that are not only more robust but also better aligned with their intended scientific objectives.

The results of our inter-individual disagreement analyses build on the well-established finding that observers can disagree on the classification or exact temporal delineation of a behaviour [21, 22]. While rates of inter-individual disagreement do not translate directly to rates of mislabels, they show that errors in labelling are in many cases unavoidable. Some steps can be taken to improve inter-observer agreement, including extensive training, more clearly defined ethograms, shorter observation sessions to avoid fatigue and recruiting observers with long-term experience observing the animals in question [23, 24]. Nevertheless, some cases are by their very nature difficult to assign, such as subtle [23], variably executed [25] or transitionary [26, 27] behaviours. When evaluating ML models, researchers should bear in mind the potential for mislabels, as these can penalise the reported performance metrics even if the model performs well (Box 1). Reviewing the confusion matrix and evaluating examples of incorrect predictions on the validation set can help identify any patterns in errors. For instance, errors may occur more often between transitional behaviours such as shifting from walking to running. Understanding the limitations of our human-labelled validation dataset (particularly when it comes to 'noisy' data) can help to set more realistic expectations for evaluation metrics and can inform decisions about when to stop training, which can avoid endlessly testing alternative models and algorithms.

In most cases, models will perform worse than hoped for on metric tests. One common pattern is that model performance is lower when there are more categories to classify (this is expected). In such cases, rather than testing the applicability of the 'imperfect' model (as presented here in our simulations), researchers are often tempted to merge categories that are frequently confused by the model and/or have similar biological functions (Ladds *et al.* 2017; Christensen *et al.* 2023) as a way of obtaining higher

values in the evaluation metrics. We note that having fewer categories will almost always increase model performance across all metrics [13]. The reason is mathematical: a completely naïve model will have a higher probability of making a correct guess if there are fewer categories to guess from. However, merging categories does not necessarily mean that the new model is more useful. For example, combining two foraging behaviours that are often confused—such as browsing a bush and pecking at ground level—might reduce the reported error, but if these behaviours are still mistaken with other non-foraging-related behaviours that the model can predict (e.g. locomoting), then merging these behaviours might not add to any conclusion about foraging in relation to the non-foraging behaviours. Instead, increasing the reported model performance may even narrow the model's applicability and hinder hypothesis testing. For instance, merging foraging-related categories could limit our ability to test individual foraging mode/diet preferences (Toscano *et al.* 2016; Gharnit *et al.* 2022) or seasonal variation in foraging modes (Owen-Smith, Fryxell & Merrill 2010). The results from our simulations provide researchers with an additional pathway to convincingly demonstrate the applicability of their models without jeopardizing the initial research questions of interest (e.g. by losing resolution by merging categories).

High performance metrics might also be misleading and lead to overconfidence in the model. Beyond the common pitfalls that might lead to overfitting models (e.g. having the same or correlated data both in training and validation datasets; Ferreira *et al.* 2020b; Aulsebrook, Jacques-Hamilton & Kempenaers 2024), performance metrics do not always capture the different contexts in which categories might occur. For example, a binary classification model that is 99% accurate (evaluated on a balanced validation dataset) in detecting the presence vs absence of birds in a frame of a video might still produce many more false positives than actual correct predictions for the less common category (in this example, presence of birds) if the number of frames without birds is hundreds of times more common than frames with birds during the real application of the model. We argue that estimating the ability to detect effect sizes and validating models by using biological hypotheses derived using *a priori* knowledge is a more reliable method to evaluate model performance because it more closely replicates how the model will be used in the future. Therefore, validating models through simulations and testing “simple” hypotheses should not be seen as last resort

to validate apparent low performing models, but as a way of validating any model regardless of their associated performance metrics.

Finally, our proposed validation approaches should not be used as a loophole to excuse poorly trained models. In the same way that reported traditional evaluation metrics can be high for low performing models (e.g. overfitted models: Bejani & Ghatee 2021), poorly trained models could output data that can appear to detect expected effect directions but totally misrepresent the effect size. The true value of validating models through simulations and hypotheses testing lies on pushing researchers to openly and critically think about the potential biases of their models and consider all possible scenarios for which their models could produce misleading scientific outcomes (see also Box 2), instead of evaluating them shallowly. We hope that these additional quality control layers can improve the ability for researchers to switch from model refinement to hypothesis testing while increasing the quality of ML-based proof-of-concepts works and applications. This can help mitigate the proliferation of publications featuring overfitted models, a trend already observed in other research fields (Navarro *et al.* 2021; Berisha *et al.* 2022; Saidi, Dasarathy & Berisha 2024). Biological validation adds a final layer of quality control by simulating real-world applications of the models before their publication and/or before using them to address real biological questions of interest. Such final layer of validation should elevate the standards of ML-related publication in ecology and evolution.

Box 1 – Effect of mislabels on training and evaluating models.

Errors in labelling training data can have different impacts on the real and estimated performance of a model. Here we introduce three types of errors into empirical data comprising classification of three species of birds: nuthatches, great tits, blue tits. Pictures were collected at artificial feeding stations and pre-processed (e.g. removing the background to select only the pixels corresponding to the birds) using the methodology described in Ferreira *et al.* (2020b). The three types of error consist of different types of erroneous labels, which are used for both training and validation dataset (see supplementary material for Box 1):

Uniform error: All pictures from all three species of birds have an equal chance of being mislabelled. This illustrates a situation in which all classes (i.e. species of birds in this example) are equally difficult to label.

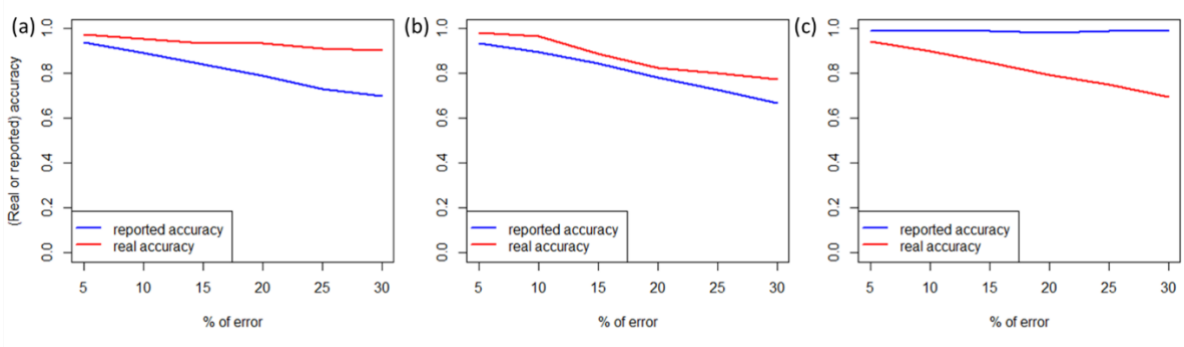
Class-dependent error: Pictures from nuthatches are always correctly labelled, but pictures of great tits could be mislabelled as blue tits and *vice-versa*. This error type represents a situation where some classes are easy to classify, but others are more difficult to distinguish.

Class- and feature-dependent error: Here we artificially introduce a feature to the pictures and correlate this feature with the probability of mislabel. For all pictures of great tits and blue tits that had a wrong label we also blurred those images. This would simulate a situation in which the human labeller would have great difficulties in classifying these two species of birds but only when the image quality was low (in this case blur).

For all types of error, we train several models (a VGG19 backbone model pre-trained on ImageNet dataset, see Supplementary material for details) that vary in the proportion of mislabels (equal for training and validation dataset) and evaluate each model using simple accuracy (N instances of predicted labels matching the annotated label/ all instances) on the validation dataset (reported accuracy). We then compare the performance to the same set of validation images but without any error in the labels, representing the true performance of the models (real accuracy).

For all cases we can see that the reported accuracy is misleading when error is present in the labels. For both uniform (a) and class-dependent error (b), the model is able to learn beyond human capacity (i.e. makes fewer errors in predicting labels than the proportion of mislabels in the training data). In this scenario, which is likely to be common in ecological research, standard metrics would underestimate the model's true performance—and this underestimation increases linearly with the extent of the error. For the class- and feature-base error (c), we would overestimate the model's ability to generalize to unseen data. This is likely because the model learns to associate the interaction between blur and pictures of great tits or blue tits, classifying all blurred pictures with the label of the other species. Since this error is also present in the validation dataset the model's accuracy is always reported as being high while the true generalization capability of the model keeps degrading. This result suggests

that it may be better practice to avoid including data that are ambiguous in the validation dataset. Finally, it is important to note that real-world error is often complex and all types of errors as well as other variations to the ones mentioned here can be present in a single dataset.



Box 2 – The impact of model error depends on the research goal.

The impact of errors in the predictions by ML-based models is not equal across all research questions. Some questions will be naturally more robust to errors, while others will be more sensitive. Imagine a ML model that is used to predict the identity of individuals and has a rate of erroneous identification (assigning the wrong identity to a detected individual). If the aim of our study was to test whether the social network of individuals is assorted by sex (Farine 2014), then even a model that makes frequent errors would still allow us to test the hypothesis robustly because assortment is very robust to imperfect sampling (Kaur *et al.* 2024) (i.e. the vast majority of edges would still accurately contribute to the test of the hypothesis). By contrast, if we wanted to instead study the spread of a contagious disease in the population, then our hypothesis may be substantially more sensitive to errors. This is because identification errors rewire the social network (Davis, Crofoot & Farine 2018). Artificially increasing the density of edges because misclassifying individual identities can introduce connections between individuals that were never (or could never have been) observed together. The addition of just a few edges could then change the properties of spread through a social network. Thus, an error rate that could yield accurate estimations for the test of one hypothesis (network assortment) could also lead to errors for another hypothesis (e.g. dramatically overestimating the potential rate of disease transmission in the population).

References

Aulsebrook, A.E., Jacques-Hamilton, R. & Kempenaers, B. (2024) Quantifying mating behaviour using accelerometry and machine learning: challenges and opportunities. *Animal Behaviour,* **207,** 55-76.

Bejani, M.M. & Ghatee, M. (2021) A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review,* **54,** 6391-6438.

Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S. & Liss, J. (2022) Are reported accuracies in the clinical speech machine learning literature overoptimistic? *Interspeech*, pp. 2453-2457.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R. & Bao, M. (2022) The values encoded in machine learning research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 173-184.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution,* **24,** 127-135.

Brewster, L., Dale, J., Guttridge, T., Gruber, S., Hansell, A., Elliott, M., Cowx, I., Whitney, N. & Gleiss, A. (2018) Development and application of a machine learning algorithm for classification of elasmobranch behaviour from accelerometry data. *Marine biology,* **165,** 62.

Brown, D.D., Kays, R., Wikelski, M., Wilson, R. & Klimley, A.P. (2013) Observing the unwatchable through acceleration logging of animal behavior. *Animal Biotelemetry,* **1,** 20.

Cagnacci, F., Boitani, L., Powell, R.A. & Boyce, M.S. (2010) Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. The Royal Society.

Chakravarty, P., Cozzi, G., Ozgul, A. & Aminian, K. (2019) A novel biomechanical approach for animal behaviour recognition using accelerometers. *Methods in Ecology and Evolution,* **10,** 802-814.

Chambers, R.D., Yoder, N.C., Carson, A.B., Junge, C., Allen, D.E., Prescott, L.M., Bradley, S., Wymore, G., Lloyd, K. & Lyle, S. (2021) Deep learning classification of canine behavior using a single collar-mounted accelerometer: Real-world validation. *Animals,* **11,** 1549.

Christensen, C., Bracken, A.M., O'Riain, M.J., Fehlmann, G., Holton, M., Hopkins, P., King, A.J. & Fürtbauer, I. (2023) Quantifying allo-grooming in wild chacma baboons (Papio ursinus) using tri-axial acceleration data and machine learning. *Royal Society open science,* **10,** 221103.

Colquhoun, D. (2017) The reproducibility of research and the misinterpretation of p-values. *Royal Society open science,* **4,** 171085.

Davis, G.H., Crofoot, M.C. & Farine, D.R. (2018) Estimating the robustness and uncertainty of animal social networks using different observational methods. *Animal Behaviour,* **141,** 29-44.

Davis, G.H., Crofoot, M.C. & Farine, D.R. (2022) Using optimal foraging theory to infer how groups make collective decisions. *Trends in Ecology & Evolution*.

Ellison, A.M. (2004) Bayesian inference in ecology. *Ecology Letters,* **7,** 509-520.

Farine, D.R. (2014) Measuring phenotypic assortment in animal social networks: weighted associations are more robust than binary edges. *Animal Behaviour,* **89,** 141-153.

Fehlmann, G., O'Riain, M.J., Hopkins, P.W., O'Sullivan, J., Holton, M.D., Shepard, E.L. & King, A.J. (2017) Identification of behaviours from accelerometer data in a wild social primate. *Animal Biotelemetry,* **5,** 6.

Ferdinandy, B., Gerencsér, L., Corrieri, L., Perez, P., Újváry, D., Csizmadia, G. & Miklósi, Á. (2020) Challenges of machine learning model validation using correlated behaviour data: evaluation of cross-validation strategies and accuracy measures. *PloS one,* **15,** e0236092.

Ferreira, A.C., Covas, R., Silva, L.R., Esteves, S.C., Duarte, I.F., Fortuna, R., Theron, F., Doutrelant, C. & Farine, D.R. (2020a) How to make methodological decisions when inferring social networks. *Ecology and Evolution,* **10,** 9132-9143.

Ferreira, A.C., Silva, L.R., Renna, F., Brandl, H.B., Renoult, J.P., Farine, D.R., Covas, R. & Doutrelant, C. (2020b) Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution,* **11,** 1072-1085.

Gharnit, E., Dammhahn, M., Garant, D. & Réale, D. (2022) Resource availability, sex, and individual differences in exploration drive individual diet specialization. *The American Naturalist,* **200,** 1-16.

Giese, L., Melzheimer, J., Bockmühl, D., Wasiolka, B., Rast, W., Berger, A. & Wachter, B. (2021) Using machine learning for remote behaviour classification—Verifying acceleration data to infer feeding events in free-ranging cheetahs. *Sensors,* **21,** 5426.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. & Altman, D.G. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology,* **31,** 337-350.

Hilborn, R. & Mangel, M. (2013) *The ecological detective: confronting models with data (MPB-28)*. Princeton University Press.

Houlahan, J.E., McKinney, S.T., Anderson, T.M. & McGill, B.J. (2017) The priority of prediction in ecological understanding. *Oikos,* **126,** 1-7.

Hounslow, J., Brewster, L., Lear, K., Guttridge, T., Daly, R., Whitney, N. & Gleiss, A. (2019) Assessing the effects of sampling frequency on behavioural classification of accelerometer data. *Journal of experimental marine biology and ecology,* **512,** 22-30.

Jennions, M.D. & Moeller, A.P. (2002) Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill'method. *Biological reviews,* **77,** 211-222.

Kaur, P., Ciuti, S., Salter-Townshend, M. & Farine, D.R. (2024) Using an agent based model to inform sampling design for animal social network analysis. *bioRxiv***,** 2024.2005. 2025.595870.

Kays, R., Crofoot, M.C., Jetz, W. & Wikelski, M. (2015) Terrestrial animal tracking as an eye on life and planet. *Science,* **348,** aaa2478.

Ladds, M.A., Thompson, A.P., Kadar, J.-P., J Slip, D., P Hocking, D. & G Harcourt, R. (2017) Super machine learning: improving accuracy and reducing variance of behaviour classification from accelerometry. *Animal Biotelemetry,* **5,** 1-9.

Navarro, C.L.A., Damen, J.A., Takada, T., Nijman, S.W., Dhiman, P., Ma, J., Collins, G.S., Bajpai, R., Riley, R.D. & Moons, K.G. (2021) Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *bmj,* **375**.

Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N. & Clune, J. (2021) A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution,* **12,** 150-161.

Owen-Smith, N., Fryxell, J.M. & Merrill, E.H. (2010) Foraging theory upscaled: the behavioural ecology of herbivore movement. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **365,** 2267-2278.

Papageorgiou, D., Cherono, W., Gall, G., Nyaguthii, B. & Farine, D.R. (2024) Testing the information centre hypothesis in a multilevel society. *Journal of Animal Ecology,* **93,** 1147-1159.

Papageorgiou, D., Rozen-Rechels, D., Nyaguthii, B. & Farine, D.R. (2021) Seasonality impacts collective movements in a wild group-living bird. *Movement ecology,* **9,** 1-12.

Pei, K., Cao, Y., Yang, J. & Jana, S. (2017) Deepxplore: Automated whitebox testing of deep learning systems. *proceedings of the 26th Symposium on Operating Systems Principles*, pp. 1-18.

Probst, P., Boulesteix, A.-L. & Bischl, B. (2019) Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research,* **20,** 1-32.

Rainio, O., Teuho, J. & Klén, R. (2024) Evaluation metrics and statistical tests for machine learning. *Scientific Reports,* **14,** 6086.

Reyes-Ortiz, J.-L., Oneto, L., Samà, A., Parra, X. & Anguita, D. (2016) Transition-aware human activity recognition using smartphones. *Neurocomputing,* **171,** 754-767.

Riaboff, L., Shalloo, L., Smeaton, A.F., Couvreur, S., Madouasse, A. & Keane, M.T. (2022) Predicting livestock behaviour using accelerometers: A systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. *Computers and electronics in agriculture,* **192,** 106610.

Rolnick, D., Veit, A., Belongie, S. & Shavit, N. (2017) Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Ryder, T.B., Horton, B.M., van den Tillaart, M., Morales, J.D.D. & Moore, I.T. (2012) Proximity data-loggers increase the quantity and quality of social network data. *Biology Letters,* **8,** 917-920.

Saidi, P., Dasarathy, G. & Berisha, V. (2024) Unraveling overoptimism and publication bias in ML-driven science. *arXiv preprint arXiv:2405.14422*.

Santema, P. & Clutton-Brock, T. (2013) Meerkat helpers increase sentinel behaviour and bipedal vigilance in the presence of pups. *Animal Behaviour,* **85,** 655-661.

Sheng, V.S., Provost, F. & Ipeirotis, P.G. (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 614-622.

Shuert, C.R., Pomeroy, P.P. & Twiss, S.D. (2018) Assessing the utility and limitations of accelerometers and machine learning approaches in classifying behaviour during lactation in a phocid seal. *Animal Biotelemetry,* **6,** 1-17.

Sipper, M. (2022) High per parameter: A large-scale study of hyperparameter tuning for machine learning Algorithms. *Algorithms,* **15,** 315.

Toscano, B.J., Gownaris, N.J., Heerhartz, S.M. & Monaco, C.J. (2016) Personality, foraging behavior and specialization: integrating behavioral and food web ecology at the individual level. *Oecologia,* **182,** 55-69.

Wang, Y., Nickel, B., Rutishauser, M., Bryce, C.M., Williams, T.M., Elkaim, G. & Wilmers, C.C. (2015) Movement, resting, and attack behaviors of wild pumas are revealed by tri-axial accelerometer measurements. *Movement ecology,* **3,** 2.

Whytock, R.C., Świeżewski, J., Zwerts, J.A., Bara-Słupski, T., Koumba Pambo, A.F., Rogala, M., Bahaa-el-din, L., Boekee, K., Brittain, S. & Cardoso, A.W. (2021) Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution,* **12,** 1080-1092.

Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M. & Fortson, L. (2019) Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution,* **10,** 80-91.

Yan, Y., Rosales, R., Fung, G., Subramanian, R. & Dy, J. (2014) Learning from multiple annotators with varying expertise. *Machine learning,* **95,** 291-327.

Young, D.S. (2018) *Handbook of regression methods*. Chapman and Hall/CRC.

Zhong, M., Taylor, R., Bates, N., Christey, D., Basnet, H., Flippin, J., Palkovitz, S., Dodhia, R. & Ferres, J.L. (2021) Acoustic detection of regionally rare bird species through deep convolutional neural networks. *Ecological Informatics,* **64,** 101333.