From raw footage to behaviour: an automated pipeline for activity monitoring and behaviour identification in the wild

Liliana R. Silva [a,b,c*] , André C. Ferreira [a,b,c,d] , Claire Doutrelant [d,e] and Rita Covas [a,b,e]

[a] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[b] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

[c] Department of Evolutionary Biology and Environmental Science, University of Zurich, Zurich 8057, Switzerland

[d] CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

[e] FitzPatrick Institute of African Ornithology, DST-NRF Centre of Excellence, University of Cape Town, Rondebosch 7701, South Africa

* Correspondence:

Liliana R. Silva

Email: lilianasilva@cibio.up

ABSTRACT

1. Studies of animal behaviour usually rely on direct observations or manual annotations of video recordings for data collection. However, such methods can be very time-consuming and inefficient, leading to sub-optimal sample sizes. Recent advances in deep-learning have shown great potential to overcome such limitations. Nonetheless, for behavioural recognition, solutions currently available are mostly focused on laboratory and/or standardized settings.
2. Here, we present a pipeline to guide researchers in building their own behavioural recognition system from their video data. Our pipeline is replicable and especially useful for studies in the wild. We take advantage of over 12215hs of manually annotated behaviours of sociable weavers (*Philetarius socius*). As behaviour classifiers we used long-short term neural networks (LSTMs) known to be suitable to classify temporal sequences of data such as video-based behavioural data.
3. LSTMs allowed us to: 1) monitor nest activity by detecting the birds' presence and simultaneously classifying the type of trajectory: i.e., nest-chamber entrance or exit; and 2) identify the respective behaviour performed: provisioning, building or aggression. With our pipeline we reduced the behavioural annotation error by 21%, while increasing by eight times the analyses speed.
4. This work demonstrates how previously analysed long-term data can be used to develop highly-efficient automated behavioural analyses, whereas this pipeline is easily adaptable to other species and behaviours, requiring only labelled video data.

Keywords

Aggression, automation, behavioural analysis, bird detection, building, deep-learning, nest provisioning, recurrent neural networks

INTRODUCTION

Animal behaviour studies frequently rely on video for behavioural characterization and quantification. However, behaviour analysis through video is increasingly costly given the complexity and variable nature of behavioural data and the ever-growing size of longitudinal behavioural datasets. On one hand, video analysis requires expertise, paid annotators or commercial software, demanding financial resources and time that could be allocated to other research tasks. On the other hand, manually coding behaviour can be subjective and error prone (Anderson & Perona, 2014), being dependent on interpretation, on attention and detection capacity of the annotator that, in turn, can be susceptible to fatigue and other factors affecting attention-span. Despite these drawbacks, video still remains one of the most effective tools for behavioural data-collection, providing reliable information while having minimal impact on animal welfare.

There is an increasing effort in automation of behavioural annotation through the application of machine learning algorithms to human and animal behavioural analysis, and more recently with particular emphasis on the use of deep-learning (DL) techniques (Christin et al., 2019). These DL models allow for automatic detection and extraction of features from complex high-dimensional data, such as behaviours, often surpassing human capabilities. During the past decade most advances on behavioural automation, either open-source or commercial hardware/software, are focused in laboratory and constricted settings for studying small animals (reviewed in Panadeiro et al., 2021; but see e.g., Mounir et al., 2023 for some examples in wild). Although they constitute important advances, such frameworks lack applicability for contexts characterized by non-controlled illumination, mutable environment and unrestricted animal movement, such as seen in wild.

Automated video-based behavioural analyses usually encompass the following steps. First, animals can be detected from the background, from which their changes in position and pose can be measured sequentially, tracking their movement through time (e.g., Weinstein, 2018). Secondly, these temporal series of selected images with associated coordinates of the animal's body are used to estimate other important behavioural descriptors such as acceleration, posture, shape, proximity that will allow further

categorization (e.g., Marks et al., 2022). However, researchers can take advantage of DL to directly classify behavioural sequences from raw video. Such approach releases the need for researchers to select specific key points/regions of interest or to calculate specific behavioural descriptors of the animals, as DL models are normally able to identify such features on their own. For example, Williams & DeLeon (2020) developed a DL-based classifier for incubation behaviour of purple martin (*Progne subis*) at nest-boxes, by providing "incubating" or "not-incubating" examples. As this strategy only requires the labelled temporal sequential examples of the behaviour(s) of interest, the classification principle is readily applicable to any dataset and system, allowing the classification of more complex behaviours that would be hard to define using solely tracking measurements.

Currently, DL is largely accessible, making it easy to demonstrate proof-of-concept ideas, leading to what can arguably be considered as over-optimism regarding application (Saidi et al., 2024), as the actual deployment of DL techniques are challenging, especially in wild settings. However, these challenges receive less attention. Moreover, model generalisation is crucial for a realistic utilisation by long-term projects due to constant adaptation over time to an ever-changing environment.

Here we propose a pipeline, from development to deployment, to guide researchers in building their own behavioural recognition system directly from their video data, and using open-source python libraries (Fig.1A). We developed this pipeline using a long-term research program that uses the sociable weaver (*Philetairus socius*) as a model species to study the evolution and consequences of sociality, and which relies on complex behavioural data collected on cooperative and agonistic behaviours, through thousands of hours of video manually analysed since 2014. Furthermore, as behaviour classifiers, we used a recurrent neural network - the long short-term memory network ("LSTM") - known to be an adequate architecture to classify temporal streams of data (Hochreiter & Schmidhuber, 1997), such as the sequence of images that compose behaviours. These networks allowed us to automatically and simultaneously detect the birds' presence and respective behaviour performed — nestling provisioning (as nest-chamber entrance or exit; hereafter nest activity), building or aggression—using only the labelled videos. We evaluated this approach by: 1) assessing the performance and speed of the model for behaviour classification, 2) comparing model performance to

analysers conducting manual annotations and 3) assessing the model's ability to output data that could detect real biological effects. Finally, we discuss some of the caveats of moving from proof-of-concept to model deployment.

METHODS

Study system and behavioural data

The sociable weaver is a colonial cooperatively breeding bird from southern Africa, living in large communal nests composed of several independent nest-chambers used for breeding. Parents are assisted by non-breeding individuals (called helpers) that help raising the young by sharing all parental care activities: provisioning, nest-chamber maintenance (through building) and nest-chamber defence (by attacking possible intruders). This results in variable family groups (2-12 individuals) with significant variation in nest-chamber visits and behaviours (from <1 to 67 visits/h). Video recordings (usually 2hs long) were collected in the wild, in a standardized way, using cameras that recorded in Full HD-quality, usually at 50 frames per second. The cameras were aimed at each nest-chamber (n=1239), and videos were manually annotated from 2014-2021. Taking advantage of this massive repertoire (5143 videos, 12215 recorded hours, 208548 nest-chamber visits), we extracted temporal-preserved frames from the manually annotated raw videos and constructed separate datasets for each type of behavioural classification task (i.e., nest activity, building and aggression), with the addition of an auxiliar smaller dataset for pre-processing to convert our annotations by second to meaningful frames (i.e., bird presence). All details for data collection, manual annotation, frames' extraction and pre-processing are in Supplementary Material S1-S4, and illustrated on Fig.1A).

Behavioural classification

For this pipeline we were interested in identifying four different behaviours (Fig.1B-1C): going inside ("IN") and outside the nest-chamber ("OUT"), as our nest-provisioning proxy, bringing a straw to build at the nest-chamber ("building") and aggressing a conspecific ejecting it outside the nest-chamber

("aggression"). The four behaviours of interest only make a minuscule part of recording time - on average a two-hour video only contains 0.23% of activity. The remaining time is hereafter referred as "negative class" ("NC", Fig. 1C). The first two behaviours make the majority of frames, while the other two occur more rarely (6.5% and 0.16% of all events are building or aggression, respectively). This results in an unbalanced data, which is problematic for DL training even when counter-acted (e.g., data augmentation). Additionally, building and aggression are dependent on nest activity behaviours. Building is associated with birds flying towards the nest-chamber ("IN") carrying a straw in the bill, while aggression is always followed by an "OUT" event of the aggressed bird (and sometimes also the aggressor(s)). Therefore, we adopted a hierarchal pipeline of three different models for classifying: i) nest activity (INs vs OUTs vs everything else), ii) building (INs with straw vs INs without straw) and iii) aggression (OUTs with aggression vs OUTs without). The idea is that first the model is used to screen the entire videos for frames that corresponded to nest activity, followed by the other two models that are only used to predict behaviours previously identified as INs (for the building model) or OUTs (for the aggression model).

Since these behaviours are hardly distinguished by humans from one single frame, we used LSTMs to classify each behaviour (Table 1). Unlike standard feedforward neural networks, recurrent networks allow a better classification of sequences by having feedback connections and "memory" of previous decisions (Hochreiter & Schmidhuber, 1997). Our models were composed of a pre-trained VGG19 backbone (without its final classification layer) for feature extraction from the images. Temporal dependencies across the frames' extracted features were captured using an LSTM layer. Afterward, four dense layers reduced the feature space. The final classification was done with a softmax layer with three output categories ("IN", "OUT" or NC) for the nest activity model and for the other two models a sigmoid layer (with one output building or NC; aggression or NC) was used for classification.

For all models we used a transfer learning approach by starting model training from a pre-trained model. For the nest activity, the VGG19 was pre-trained on ImageNet dataset. For building and aggression detection, we used the nest activity model as a starting training point. We did this since the activity

model's dataset was more similar to the building and aggression classifications than other commonly used pre-trained models (e.g., using ImageNet dataset).

For all the three models we used as input a sequence of six frames. Frames were consecutive for the nest activity and building models. However, for the aggression model, we used a sequence of six frames, skipping every two frames (i.e., the first and last frames of the six-frame sequence are 16 frames apart) as observers usually needed more frames (or frames further apart) to recognize aggressions.

The nest activity model was trained on the largest dataset (Table 1), being crucial to prevent misclassifications by identifying only the periods of activity for which the other two behaviours could occur. Moreover, in order to minimize the rate of nest activity false positives (i.e., predicting behaviour when it does not exist) we carefully built a "negative class" dataset instead of randomly sample from the videos. This class includes sequences of frames that are potentially mistaken with INS and OUTs (Fig.1D), as randomly sampling from the data would have many "easy" to classify frames (i.e., static nest-chamber-entrance). For building and aggression, we also carefully selected "hard" negative examples instead of random sampling from the videos.

See Supplementary Material S5 and scripts for additional details on the training process.

Evaluation

We first evaluated our models' accuracy using a balanced validation dataset (i.e., equal number of instances of each class) (Table 1). Nonetheless, to assess true model applicability and generalization for potential deployment, all models were compared to manual annotations of full video recordings never utilized during model training (and only used once for final evaluation). We assessed models' production of false positives (i.e., when a behaviour is annotated for no relevant activity; "FP") and false negatives (i.e., when a behaviour of interest was not detected; "FN"). Additionally, model performance was compared to human analysis capacities in terms of percentage of FP (for each video: number of false positives/number of real visits) and FN (number of false negatives/number real of visits), as well as speed (Table 1). Finally, nest activity and building models were biologically validated

by testing the models' ability to output predictions that could be used to test hypotheses with *a priori* known outcomes: 1) that age and number of nestlings is positively related with nest activity and 2) building activity is related with the nest stage (higher during incubation) (Table 1). For aggression, however, no a priori predictions were available. All details for models' evaluation and video processing are in Supplementary Material S6.

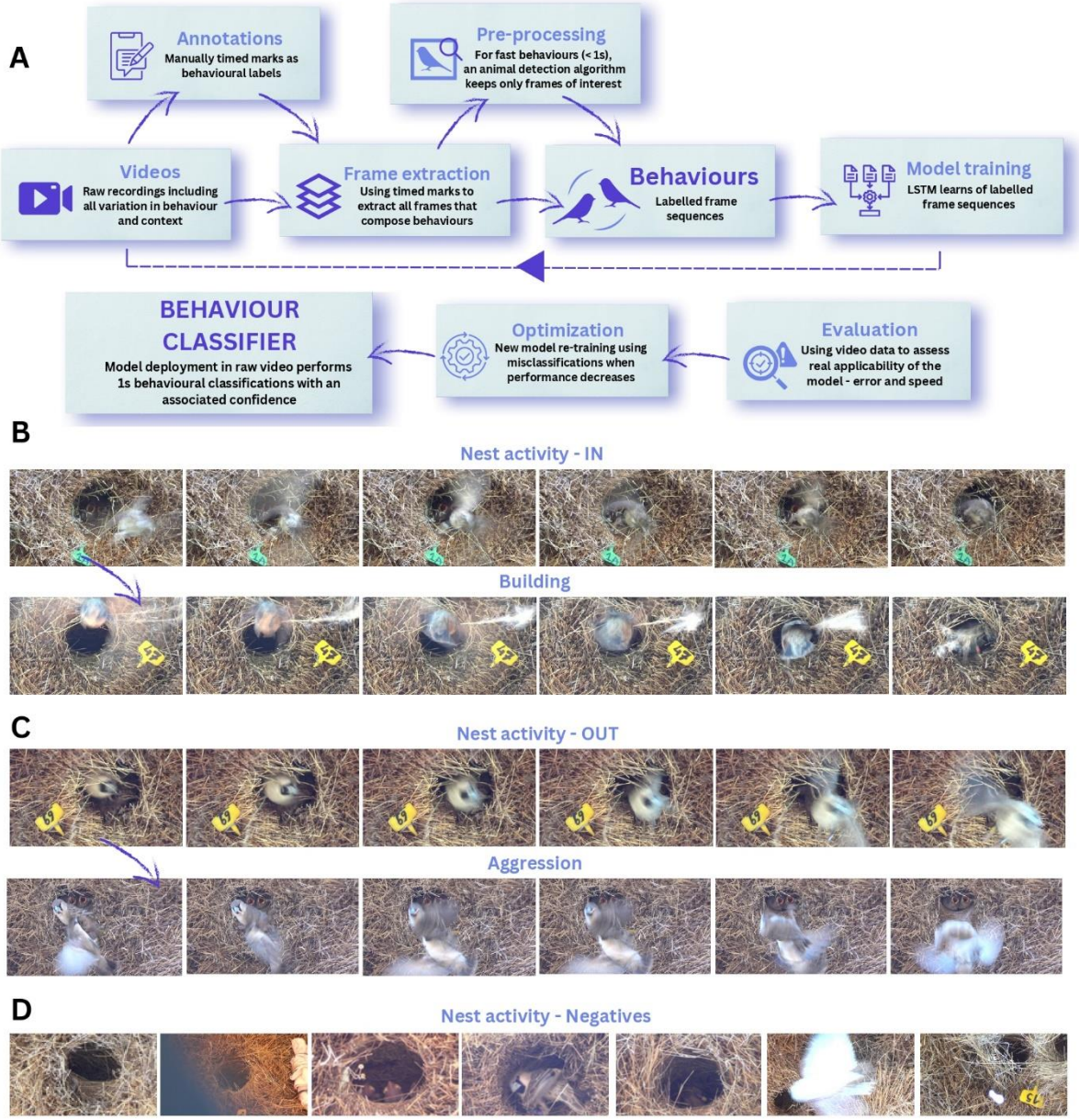

Figure 1: A) Pipeline for automatic behavioural classification from video by using manually annotated data for the application of LSTMs. Frame sequence examples depicting the hierarchical automation of behaviours in the sociable weaver: B) “IN” when a bird enters and “building” when it brings a straw, whereas C) “OUT” when a bird leaves and “aggression” when a bird(s) expels another from the nest-chamber. D) Single frame examples of negative class provided to the nest activity model to decrease possible false positives (i.e., predicting behaviour when it does not exist). Example frames shown from left to right: empty nest-chamber, microphone bag and light obstruction, chicks, perching bird outside and inside the nest chamber, passing bird not visiting the focus chamber, a faecal sac ejection.

Table 1: Behavioural detection models' training details, model performance and speed compared to observers. Model training details include: classes, interval of frames, training and validation datasets sizes. Examples of negative class ("NC") allows to include variation in the models. For more effective use, model processing was hierarchical and dependent on behaviour type: nest activity model allows to detect timestamps for the other two models (making them faster). Model performance was assessed as the model frame-accuracy obtained at the validation stage, but also, in terms of 2h videos (our unit of analysis) by assessing the produced false positives ("FP", i.e., when a behaviour is annotated for no relevant activity) and false negatives ("FN", i.e., when a behaviour of interest was not annotated), as well as speed performed by model and observers.

[a] Nest-chamber entrance, [b] nest-chamber exit, * every three frames, ‡ including individual identification.

Model	Classes	Frames interval	Frames training	Frames validation	NC composition	Processing	Model accuracy	Model FP | FN	Observers FP | FN	Model speed	Observer speed
Nest activity	IN[a] vs. OUT[b] vs. NC	6	371775	59484	empty nest, chicks, sanitation, objects, perched/passing birds, light	full video	95.7 %	6.24 % 0.39 %	10.05 % 3.36 %	28.6 min	40.45 min
Building	building vs. NC	6	78732	25444	INs, bringing prey	only INs	87.5%	3.97 % 2.3 %	0.58% 4.97%	0.51 min	17.73‡ min
Aggression	aggression vs. NC	6*	9144	1152	OUTs	only OUTs	88.02%	5.82 % 0.06 %	1.62% 3.22%	0.05 min	17.73‡ min

RESULTS

All models had a validation accuracy between 87.5-95.7% (Table 1) and error rates were equivalent or superior to annotators (Table 1). All models greatly surpassed annotators in analysis speed (Table 1). Both the nest activity and building models allowed for the detection of biologically relevant effects (see Fig.2-3).

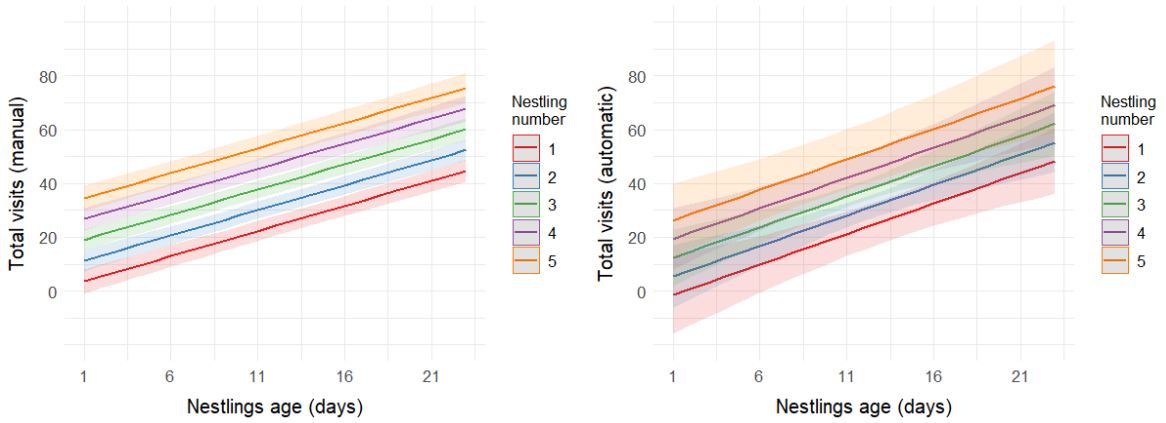

Figure 2: Predicted total visits to the nest with increasing nestling age, number of nestlings in the nest and video length for (A) manual and (B) model annotations, using linear models. Solid lines represent model predictions, colours represent nestling count category, coloured shadows represent model confidence intervals. Modelling details in Supplementary Material S6.

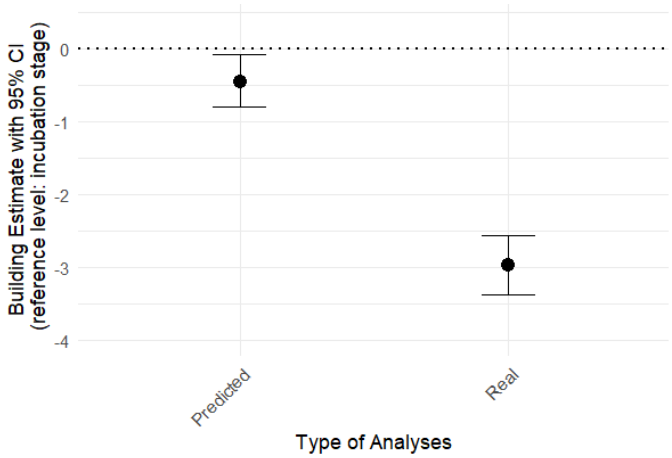

Figure 3: Estimated effect sizes (log-odds) with 95% confidence intervals (CI) from the GLM predicting the likelihood of building (yes/no) per visit, as a function of breeding stage (incubation vs. nestling).

Negative estimates with CIs that do not overlap zero indicate a reduced likelihood of building during the nestling stage. Modelling details in Supplementary Material S6.

DISCUSSION

With this work we built a powerful behavioural analysis pipeline that automatizes the detection and annotation of nest activity, building and aggression of sociable weavers. Our pipeline outperformed human annotators—both experienced and non-experienced. The higher performance of our models, coupled with the 24/7 working time of a computer, allowed us to increases by eight times the processing speed analyses (from analysing 41.25 to 345.68 videos per week), and reduce by ca. a fifth the error (in the form of wrong detections or missed behaviours, less 21.09% error, see Supplementary Material S6C).This pipeline has already been successfully used to analyse automatically over 4095 (2hs) videos of nestling provisioning, egg incubation and roosting behaviour (e.g., Fortuna et al., 2024). For nest activity alone, this automation already saved >2600 human working hours. Our results show that other researchers can use their own previously annotated behavioural data to build powerful classification models to increase the speed and efficiency of their behavioural analyses.

Automation through recurrent neural networks proved to be successful in automating behaviour. As indicated by the error from annotators, the behaviours automated in this work were not easy to classify. Nonetheless, our models produced less combined error (summing all sources of error of all models, 18.78%) than the average observer (23.8%). Additionally, it is consistently comparable to experienced annotators (>six months of analysis) and a great improvement from non-experienced annotators for most of the tasks (see Supplementary Material S6 for error discrimination annotators experience). The exception being of nest-chamber building, a behaviour that can be hard to distinguish given that the straws carried by the birds are similar to the background, for which we had a slightly larger error per visit compared to annotators. Nevertheless, the data predicted from our model still managed to detect a known biological feature on a relatively small testing dataset (higher building investment during incubation). For aggression, given the rarity of this behaviour, a biological validation was not possible,

nevertheless we showed that our model can identify aggressions in a more accurate way than annotators, missing only one event during testing.

By relying on raw labelled behavioural data, recurrent networks are a versatile way of automating behaviour. For example, in the case of the sociable weaver, our models are also useful under various circumstances such as low-light conditions (typical of roosting-time recordings), with visiting predators (e.g., gabar goshawk *Micronisus gabar*), and identifying behaviours even when several individuals are present. Furthermore, we showed the versatility of this approach by adapting a model trained on one behaviour with a larger dataset available (nest activity model) and re-trained it for detection of less common behaviours (approximately 3.6- and 28.6-times smaller datasets for building and aggression, respectively). Our approach shows that not only our pipeline is potentially applicable to other behaviours but also that our models (and data) can potentially be used for other tasks, contexts and species (e.g., activity at nest-boxes). Furthermore, although scarcely applied to video, recurrent networks have been proved to be successful at detecting behaviour in animals through wearable sensors (Mao et al., 2023) and vocalizations (e.g., Müller et al., 2021). Future work, could integrate this type of models and dimensions of behavioural data (image, sensoring and sound) to produce better behaviour descriptions.

Although extremely useful, automation of behavioural analyses through DL, especially for video, has limitations, particularly during deployment. First, model performance can decline over time affecting deployment. Specifically, factors associated to long-term studies such as material turnover (e.g., changes in camera quality), new research tools (e.g., microphones attached to the nests, in Fig.1), appearance of other species (e.g., unexpected predators), protocol adjustments (e.g., shifting recording starting times) can introduce changes in the data that were not initially included in the training and decreasing model performance. Therefore, for long-term deployments, DL models require ongoing monitoring and optimization (Fig.1A). Monitoring models' performance involves regular inspection of models' outputs, which in turn, creates an opportunity to take advantage of model mispredictions to optimize them by re-training models with "hard" examples. This process enhances the models' generalization and longevity—key objectives for researchers using DL, given the significant investment

in developing and maintaining DL approaches. Finally, developing and deploying such models requires time, powerful computers, storage resources and a high energy consumption. A balance must be struck between the need for automation, for example due to a constant accumulation and use of behavioural data, and the considerations mentioned above—whether long-term applicability is required or resources are limited.

ACKNOWLEDGEMENTS

We thank Babette Fourie, Cecile Vansteenberghe, Franck Theron, Hugo Pereira, Marina Sentis, Myriam El Harouchy, Sandra Esteves, Sophie Lardy, Rita Fortuna, Rita Leal, Zoe Tarren and several other field team managers, assistants and volunteers for help with data collection in the field. De Beers Consolidated Mines gave us permission to work at Benfontein Reserve. This study was supported by funding from ERC (EU, Consolidator grant 866489), FCT (Portugal, grants IF/01411/2014/CP1256/CT0007 and PTDC/BIA-EVF/5249/2014) and DST-NRF Centre of Excellence at the Fitzpatrick Institute of African Ornithology University of Cape Town to RC, ANR (France, grant 19-CE02-0014-01) to CD and Marie Curie-Staff Exchange (grant 101183160) to all. This project is part of the OSU-OREME, and long-term Studies in Ecology and Evolution (SEE-Life) program of the CNRS.

CONFLICT OF INTEREST STATEMENT

None of the authors have any conflicts of interest.

AUTHOR CONTRIBUTIONS

All authors conceived the idea, contributed to the methodological design, video-data collection and management. LRS performed most of the manual behavioural analysis. LRS and ACF developed the deep-learning pipeline. LRS led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

REFERENCES

Anderson, D. J., & Perona, P. (2014). Toward a Science of Computational Ethology. Neuron, 84(1), 18–31. https://doi.org/10.1016/J.NEURON.2014.09.005

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. Methods in Ecology and Evolution, 10(10), 1632–1644. https://doi.org/10.1111/2041-210X.13256

Fortuna, R., Covas, R., D'Amelio, P. B., Silva, L. R., Parenteau, C., Bliard, L., Rybak, F., Doutrelant, C., & Paquet, M. (2024). Interplay of cooperative breeding and predation risk on egg allocation and reproductive output. Behavioral Ecology, 35(2). https://doi.org/10.1093/beheco/arae010

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Mao, A., Huang, E., Wang, X., & Liu, K. (2023). Deep learning-based animal activity recognition with wearable sensors: Overview, challenges, and future directions. Computers and Electronics in Agriculture, 211, 108043. https://doi.org/10.1016/J.COMPAG.2023.108043

Marks, M., Jin, Q., Sturman, O., von Ziegler, L., Kollmorgen, S., von der Behrens, W., Mante, V., Bohacek, J., & Yanik, M. F. (2022). Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. Nature Machine Intelligence 2022 4:4, 4(4), 331–340. https://doi.org/10.1038/s42256-022-00477-5

Mounir, R., Shahabaz, A., Gula, R., Theuerkauf, J., & Sarkar, S. (2023). Towards Automated Ethogramming: Cognitively-Inspired Event Segmentation for Streaming Wildlife Video Monitoring. International Journal of Computer Vision. https://doi.org/10.1007/S11263-023-01781-2

Müller, R., Illium, S., & Linnhoff-Popien, C. (2021). A deep and recurrent architecture for primate vocalization classification. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 6. https://doi.org/10.21437/Interspeech.2021-1274

Panadeiro, V., Rodriguez, A., Henry, J., Wlodkowic, D., & Andersson, M. (2021). A review of 28 free animal-tracking software applications: current features and limitations. Lab Animal, 50(9), 246–254. https://doi.org/10.1038/S41684-021-00811-1

Saidi, P., Dasarathy, G., & Berisha, V. (2024). Unraveling overoptimism and publication bias in ML-driven science. https://arxiv.org/abs/2405.14422v3

Weinstein, B. G. (2018). Scene-specific convolutional neural networks for video-based biodiversity detection. Methods in Ecology and Evolution, 9(6), 1435–1441. https://doi.org/10.1111/2041-210X.13011

Williams, H. M., & DeLeon, R. L. (2020). Deep learning analysis of nest camera video recordings reveals temperature-sensitive incubation behavior in the purple martin (Progne subis). Behavioral Ecology and Sociobiology, 74(1), 1–12. https://doi.org/10.1007/S00265-019-2789-2/FIGURES/4