From video to behaviour: an LSTM-based approach for automated nest behaviour recognition in the wild

Running Headline: LSTM for behaviour recognition

Liliana R. Silva [a,b,c*] , André C. Ferreira [a,b,c,d], Irene Martínez-Baquero [e], Arlette Fauteux [f], Claire Doutrelant [d,g] and Rita Covas [a,b,g]

[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] Edward Grey Institute of Field Ornithology, Department of Biology, University of Oxford, United Kingdom

[f] Département des Sciences Biologiques, Université du Québec à Montréal, CP-8888 Succursale Centre-ville, Montréal, Québec, Canada

g 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. However, such methods can be very time-consuming and error-prone, leading to sub-optimal sample sizes. Recent advances in deep-learning show great potential to overcome such limitations, however, most currently available behavioural recognition solutions remain focused on standardised settings.
2. Here, we present a deployment-focused framework to guide researchers in building behavioural recognition systems from video data, using Long Short-Term Memory (LSTM) networks to classify behavioural sequences across consecutive frames.
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 behaviour performed: building, aggression or sanitation. Using our framework, we largely outperformed human annotators in error and speed. LSTM performance improved with challenging training instances, and remained robust even with modest sample sizes. LSTM also outperformed YOLO, highlighting the critical role of temporal sequence information in behavioural analysis.
4. We demonstrate that our approach is replicable across three bird species and applicable to deployment videos, highlighting its value as a generalizable and transferable tool for long-term studies in the wild.

Keywords

Behavioural analysis, bird, deep-learning, deployment, LSTM, nest behaviours, recurrent neural networks, YOLO

INTRODUCTION

Animal behaviour studies frequently rely on video for behavioural characterisation and quantification. However, behaviour analysis through video is costly given the complexity and variable nature of behavioural data and the ever-growing size of longitudinal datasets. Video analysis requires expertise, annotators, or commercial software, demanding significant resources, whereas manual coding is subjective, error-prone and tedious (Anderson & Perona, 2014). Despite these drawbacks, video still remains one of the most effective tools for behavioural data-collection.

There is increasing effort to automate behavioural annotation by applying machine learning algorithms to human and animal behavioural datasets, with recent 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, often surpassing human capabilities (Pichler & Hartig, 2023). Most recent advances on behavioural automation from video, either open-source (e.g., Harris et al., 2023) or commercial hardware/software (e.g., Ethovision, SMART), are focused on constricted settings (reviewed in Panadeiro et al., 2021; but see e.g., Mounir et al., 2023). Although they constitute important advances, such frameworks lack applicability for contexts characterised by varying recording conditions, mutable environment and unrestricted movement, such as seen in the wild.

Automated video-based behavioural analyses require distinguishing the animal(s) against the background, and their behaviour can then be inferred from either a single frame (e.g., Norouzzadeh et al., 2018) or a sequence of frames (e.g., Williams & DeLeon, 2020). Single-frame approaches risk missing the dynamic nature of behaviour, while multi-frame methods better capture changes over time (e.g., Bohnslav et al., 2021). For example, a bird flying toward a nest may appear to be entering in one frame, but only a sequence reveals whether it actually does. Multi-frame approaches are potentially more reliable by considering the sequential changes in position and pose over time (e.g., Weinstein, 2018). Temporal series of images can be analysed by associating coordinates of the animal's body to behavioural descriptors such as acceleration, posture, and shape, that are ultimately translated to a

behavioural classification (e.g., Marks et al., 2022). However, rather than relying on manually defined landmarks, deep learning models such as long short-term memory networks (LSTMs) can automatically model temporal dependencies in sequential data, making them well suited for classification tasks involving streams of data (Hochreiter & Schmidhuber, 1997). These networks allow to simultaneously detect the animals and respective behaviour using only annotated video segments, avoiding the need to select key points/regions or to calculate specific behavioural descriptors of the animals. While such approaches have the potential to revolutionise video-based animal behavioural analyses, they have been mostly applied to humans (e.g., Van Houdt et al., 2020; but see Williams & DeLeon, 2020) or other animal data types (e.g. accelerometer: Mao et al., 2023; sound: Müller et al., 2021).

Currently, DL is largely accessible, making it easy to demonstrate proof-of-concept ideas, which has arguably led to over-optimism regarding its application (Saidi et al., 2024). However, the deployment of DL models is challenging, especially for a realistic application in the wild and long-term settings as models that perform well during development might perform poorly when taken to the real-world (see Wilson et al., 2025). This can be a result of multiple factors from poorly designed training and testing datasets to the appearance of unexpected instances in the wild or insufficient understanding of deployment contexts.

Here, we present an LSTM framework developed to guide researchers in building their own behavioural recognition system directly from video data, using open-source Python libraries (Fig. 1A), with a focus on realistic, long-term applications in the wild. Our framework builds on a long-term study of the sociable weaver (*Philetairus socius*), based on complex behavioural data manually collected since 2014, and offers a basis for models that can be extended to other systems. We first describe model development, including considerations related to the selection of challenging training instances, highly unbalanced behavioural samples, variations in behaviour duration, and the impact of dataset size on LSTM learning. We then evaluate our approach by: 1) assessing the performance and speed of the model for behaviour classification, 2) comparing model performance to human analysers and 3) assessing the model's ability to output data that could detect real biological effects. In addition,

we compare our multi-frame approach with a single-frame behaviour classification based on YOLO ("You Only Look Once", e.g., Harris et al., 2023). To assess the generalisation capabilities approach, we use our framework to analyse the nesting behaviour of two other species of birds: blue (*Cyanistes caeruleus*) and great tits (*Parus major*). Finally, we discuss some of the caveats of moving from proof-of-concept to model deployment.

METHODS

Behavioural data

Between 2014 and 2021, video recordings of sociable weavers were collected in the wild near Kimberley, South Africa. Recordings were standardised, with each nest chamber filmed in Full HD quality for approximately two hours, using fixed cameras. Taking advantage of this massive repertoire (n= 1239 nests, 5143 videos, 208548 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). These behaviours were manually annotated per second by human observers, requiring a pre-processing step to keep the specific frames containing the behaviour. All details for data collection, manual annotation, frames' extraction and pre-processing are in the Supplementary Material, hereafter referred to as "SM", S1:S4, and illustrated on Fig.1A).

Behavioural classification

For behaviour classification of sociable weavers, we identified four distinct behaviours (Fig. 1B–1C): entering the nest chamber ("IN"), used as a proxy for nest provisioning; exiting the nest chamber ("OUT"); bringing straws to the nest chamber for construction ("building"); and exhibiting aggression by expelling a conspecific from the nest chamber ("aggression"). These four behaviours only make a minuscule part of all recording time (<1%). The remaining time is hereafter referred to as "negative class" ("NC", Fig. 1C). Thus, we targeted "NC" frame sequences likely to be confused with the four

behaviours (i.e., hard negatives; Fig. 1D, Fig. 2), rather than randomly sampling non-behavioural segments, which would yield mostly easy negatives (e.g., static nest chambers). To quantify their impact, we compared model performance with varying proportions of hard negatives (10% vs. 55% of NC composed of selected hard negatives; see SM-S6E for modelling details and results).

Additional challenges arise because "IN," and "OUT" comprise the vast majority of behavioural frames, while building (6.5%) and aggression (0.16%) are the least common. Moreover, aggression events are longer (≈16 frames vs. ≈6 for other behaviours), reflecting the fact that human observers typically require more frames to recognise them. To address this, and given that both building and aggression are context-dependent (building occurs only when birds enter with straw in their bills, while aggression is consistently followed by the aggressed bird leaving), we implemented a hierarchical framework consisting of three models: (i) nest activity (IN vs. OUT vs. NC), (ii) building (IN with straw vs. IN without straw), and (iii) aggression (OUT with aggression vs. OUT without aggression). The first model, trained on a larger dataset, screens the entire video for nest activity, while the latter two, trained on smaller datasets, make predictions only within events identified as IN (building) or OUT (aggression). Moreover, to further accommodate differences in behaviour duration, all three models used six-frame input sequences: frames were consecutive for the nest activity and building models, while the aggression model used six frames sampled every three frames (spanning 16 frames in total).

All models combined a pre-trained VGG19 backbone (without the final classification layer) for frame-level feature extraction with an LSTM layer to capture temporal dependencies. Features were further reduced by four dense layers, and classification was performed using a softmax layer for nest activity ("IN," "OUT," "NC") or a sigmoid layer for building (IN with straw vs. IN without) and aggression (OUT with aggression vs. OUT without). Transfer learning was applied by initialising the nest activity model with ImageNet weights (following a previous classification task in this species; Ferreira et al., 2020), and subsequently using the trained nest activity model to initialise the building and aggression models, given their closer similarity. All training details can be found on SM-S5.

Because model performance can depend on training dataset size, we evaluated the error rates of models developed for nest activity detection using different dataset proportions (25%, 55%, 75%, and 100% of the data; see SM-S6F for modelling details and results).

Finally, we compared our framework with a single-frame approach based on YOLOv8 (Ultralytics, 2023) for nest activity detection using equivalent datasets (see SM-S6G for modelling details and results).

Evaluation

We first evaluated model accuracy using balanced validation datasets (Table 1). To test real-world applicability and generalisation, models were evaluated on manually annotated full video recordings, matching those used in future deployment ("deployment videos"), and reserved solely for final testing. Performance was measured in terms of error and processing speed, compared to human annotators, with a second observer as ground truth (Table 1). Error was quantified as false positives (FP; behaviours annotated without relevant activity) and false negatives (FN; missed behaviours of interest), over the number of actual visits per video. These measures directly capture deployment implications: a 100% FP rate means all predictions are wrong, while a 100% FN rate means all real behaviours are missed. Complementary evaluation metrics are provided in the SM-6D. In addition, nest activity and building models were biologically validated by testing models' predictions that could be used to directly test hypotheses with *a priori* known outcomes: 1) that age and number of nestlings are positively related to nest activity and 2) building activity is related to the nest stage (higher during incubation). For aggression, no a priori predictions were available. All details of models' evaluation are provided in SM-S6.

Species and context generalisation

To further test the generalisability of our framework, we automated behaviour detection of two additional species, monitored in 2024, as part of two long-term studies: blue tits in Corsica (France)

and great tits in Wytham Woods (UK). In both cases, the breeding pairs were recorded in settings that were differed remarkably from the sociable weavers. These recordings were made from inside the nest box using an embedded camera positioned to record the entrance, depicting the provisioning of nestlings. Blue tit recordings were collected continuously (n=17 nest-boxes, 185 videos), capturing the activity inside the nest, including sanitation events (i.e., when a bird carries a faecal bag outside). By contrast, great tit recordings lasted 2hs (n=15 nest-boxes, 38 videos), filmed from the back, were limited mostly to distinguishing arrivals and departures. Building on the approach developed for sociable weavers, we automated the detection of “IN,” “OUT,” and “sanitation” events for these two tit species (Table 1, Fig. 2). See SM-S7:S8 for additional details on the training and evaluation processes.

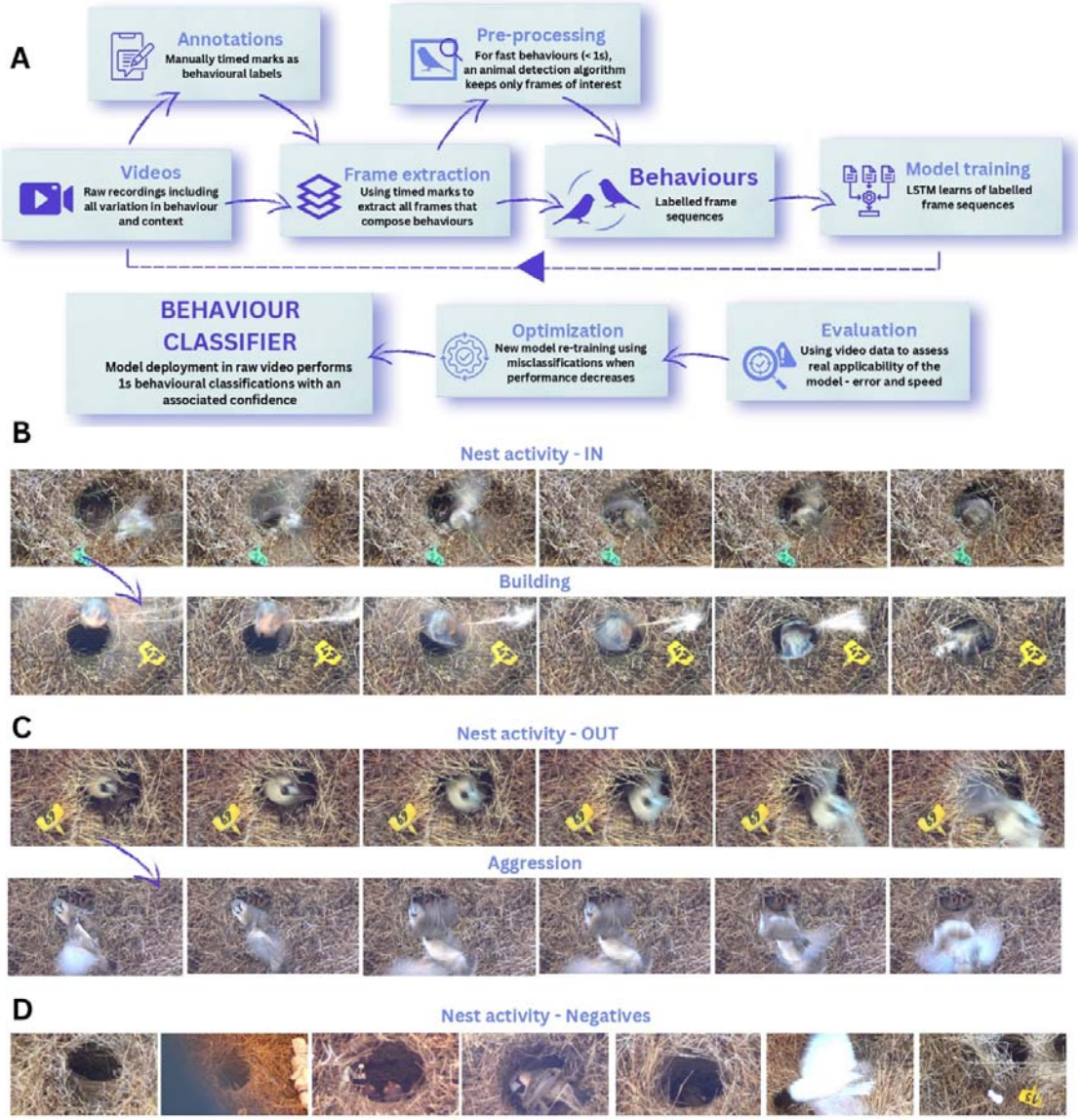


Figure 1: (A) Pipeline for automatic behavioural classification from video using manually annotated data with LSTMs. Frame sequences showing hierarchical behaviour detection in sociable weavers: (B) “IN” with “building” when entering with straw; (C) “OUT” with “aggression” when expelling another bird. (D) Single-frame negative class examples used to improve model efficiency.

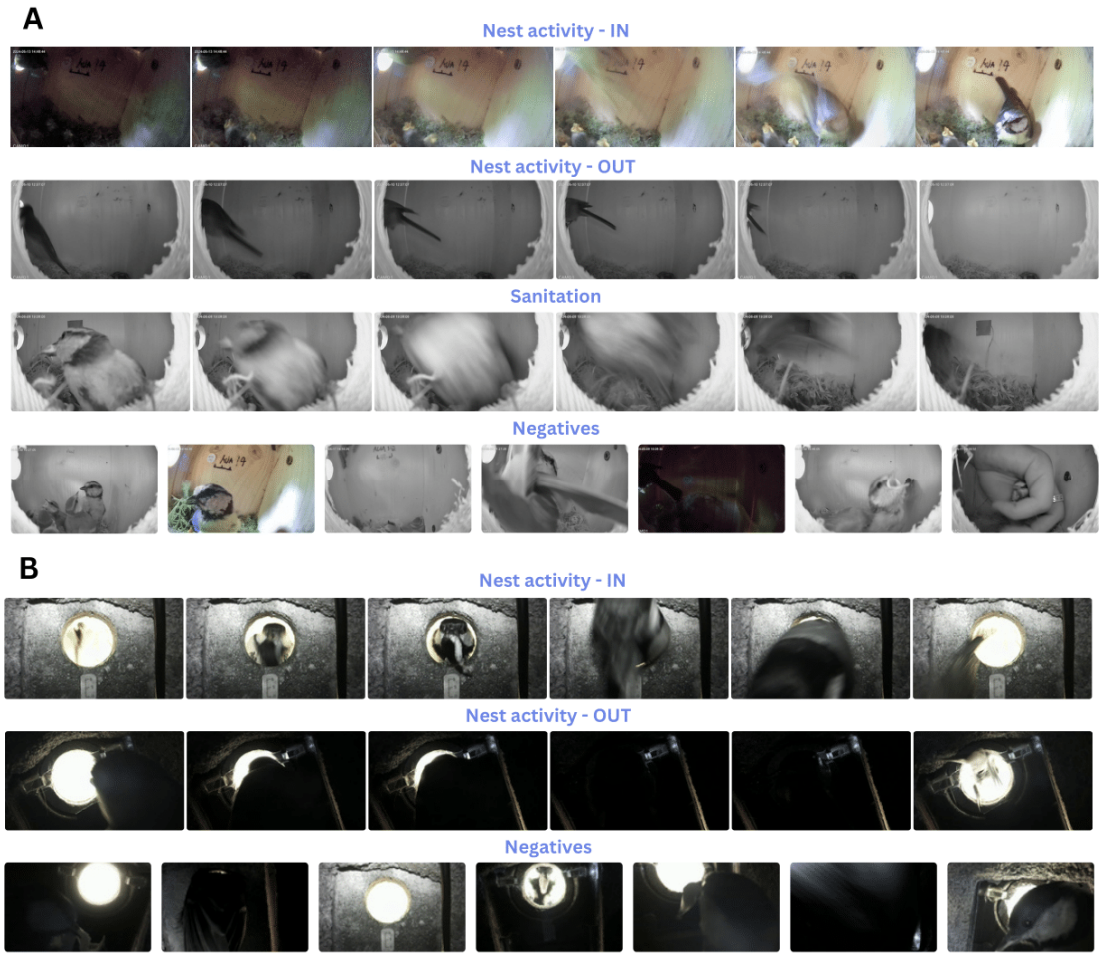


Figure 2: Frame examples of automated nest-box activity detection in (A) blue tits and (B) great tits, with corresponding negative classes. In blue tits, sanitation was hierarchically detected on exit.

RESULTS

In the sociable weaver, all models had a validation accuracy >87.5% (Table 1). FP and FN were overall lower than those of annotators, while all models outperformed them in analysis speed (Table 1). Automated data allowed for the detection of biologically significant effects, previously known for sociable weavers (details and results in SM-S6A:B, Fig.S2-S3).

Using nest activity automation in this species as an example, we found that: (1) including hard negatives during training had a strong effect, reducing error by ~60% (from 19% FP and 1.02% FN to 7.67% FP and 0.96% FN, SM-S6E); (2) overall training dataset size was not limiting, as even 25% of the original dataset (10,037 sequences) yielded error rates close to those obtained with the full dataset (SM-S6F); and (3) adopting a single-frame YOLOv8 approach markedly worsened performance, increasing error nearly fourfold (from 7.4% FP and 6.84% FN with LSTM on a comparable dataset to 54.7% FP and 1.06% FN using YOLOv8, SM-S6G).

Our framework was successfully applied to the behaviours of two tit species, also achieving high accuracy, speed, and low video error (Table 1, modelling details in SM-S7:S8).

DISCUSSION

With this work we built a powerful behavioural analysis framework that automates the identification of nest behaviours. We show that LSTMs are effective for automating behaviour classification and enable researchers to leverage their own previously annotated data, often already existent in long-term projects. While the behaviours automated in this work were challenging to classify (Table 1, “Observers FP% | FN%”), our models matched experienced annotators (with over six months of practice) and outperformed inexperienced ones (see SM-S6A:B for error rates by experience). Additionally, our framework, coupled with the 24/7 working time of a computer, allowed us to increase by eight times the processing speed analyses (from analysing 41.25 to 345.68 videos per week, see SM-S6C). The effectiveness of this framework is further demonstrated by its current

application in the sociable weaver long-term project, where it reduced manual effort by over 2,600 working hours across four breeding field seasons of video collection. Additionally, we demonstrate that the approach developed here can be readily applied to other wild birds, enabling automation that can now be implemented in other long-term studies.

Dataset quality and composition are key factors when training DL models (Gong et al., 2023). Reducing the sociable weavers' nest activity dataset to ~10,000 behavioural sequences (25% of the original size; SM-S6F), or even further to ~5,000 sequences (~12% of the original dataset, as used in the YOLO comparison analysis; SM-S6G), had little effect on LSTM performance. However, when most of the negative instances (NC) were easy classification examples, the error doubled on deployment videos, even though model train-validation metrics were similar for models with low and high hard negatives (accuracies: 97.30% and 94.44%, respectively, SM-S6E). These results highlight the need for high-quality training datasets and evaluation on deployment videos to ensure robust real-world performance. Therefore, here we provide the largest annotated dataset for nest-related behaviour to date, compiled over several years by multiple annotators and capturing wide natural variability, which is essential for robust generalisation (SM-S9). By contributing to the growing body of behavioural annotation datasets (e.g., MammalNet, Chen et al., 2023; KABR, Kholiavchenko et al., 2024) and providing a rare large-scale dataset on bird behaviour, our work facilitates cross-species comparisons, benchmarking, and transfer learning.

Comparing a single-frame approach (YOLOv8) with LSTM modelling, we found that YOLO performed substantially worse. This was expected, since behaviours are defined by actions unfolding over time (Bohnslav et al., 2021). Although both models showed promising train–validation metrics (accuracies: 96.11% for LSTM, 80.56% for YOLO; SM-S6G), the number of false positives during deployment differed drastically: from 960 real visits, YOLO predicted 5,990, whereas the LSTM model predicted 932. The LSTM slightly underestimates visits due to missed detections but remains close to the real scenario, whereas YOLO drastically overestimates by marking almost continuously and producing correct detections amid numerous false positives. Nonetheless, single-frame methods may still succeed when a specific pose or appearance uniquely indicates a behaviour, or when

combined with manually defined rules (e.g., trajectory tracking). Future work should further investigate which approach is more suitable for each context.

Although useful, automation of behavioural analyses through DL, especially for video, has limitations. First, model performance can decline over time affecting deployment. Specifically, factors associated with long-term studies, such as systematic changes in recording conditions and study design, can introduce unexpected variation not included in the training, reducing model performance. Therefore, for long-term deployments, DL models require ongoing monitoring and optimisation (Fig.1A). Monitoring models' performance involves regular inspection of models' outputs and reusing model mispredictions to re-train models with new "harder" examples. This process enhances the models' generalisation and longevity—key objectives for researchers using DL. In addition, developing and deploying such models takes time (ranging from 3-6 months depending on the complexity), highlighting the importance of creating generalisable and transferable automation approaches that can be shared across research projects, as was done here. Finally, this automation work focused on short (<1s), dynamic, non-state complex behaviours, but further research should test whether LSTMs can extend to longer behaviours.

ACKNOWLEDGEMENTS

We thank several researchers, field team managers, assistants and volunteers for help with data collection, management and analysis from the Sociable Weaver, the Corsican UQUAM-CEFE Tit and Wytham Tit long-term projects (detailed in the SM-S7 section). 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 awarded to RC, ANR (France, grants ANR-15-CE32-0012-02 and 19-CE02-0014-01) to CD and Marie Curie-Staff Exchange (Horizon MSCA grant 101183160) to LS, ACF, CD and RC. Blue tit data collection and AF were supported by the Canada Graduate Research Scholarship and Discovery Grant from Natural Sciences and Engineering Research Council of Canada (NSERC). Sociable weaver and blue tit long-term projects are part of the

OSU-OREME, and the long-term studies in Ecology and Evolution (SEE-Life) program of the CNRS. Great tit data collection and IMB were supported by the ERC (grant 250164) and the UKRI Frontiers award EP/X024520/1 awarded to Ben C. Sheldon.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

LRS, ACF, CD and RC conceived the idea and methodological design. All authors contributed to their system's video-data collection, management and analysis. LRS and ACF developed the deep-learning pipeline. LRS led the writing; all authors contributed to drafts and approved submission.

DATA AVAILABILITY

Scripts, models, and data required to reproduce this work are available via Google Drive (https://drive.google.com/drive/folders/1C0hfW8Dxc6MQnzIaWaETqp_fCHCsprL8) for peer review, and will be made publicly accessible through the Dryad Digital Repository upon acceptance.

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

Bohnslav, J. P., Wimalasena, N. K., Clausing, K. J., Dai, Y. Y., Yarmolinsky, D. A., Cruz, T., Kashlan, A. D., Chiappe, M. E., Orefice, L. L., Woolf, C. J., & Harvey, C. D. (2021). DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. eLife, 10, e63377. https://doi.org/10.7554/eLife.63377

Chen, J., Hu, M., Coker, D. J., Berumen, M. L., Costelloe, B., Beery, S., Rohrbach, A., & Elhoseiny, M. (2023). MammalNet: A large-scale video benchmark for mammal recognition and behavior understanding. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (pp. 13052–13061). https://doi.org/10.48550/arXiv.2306.00576

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

Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R., & Doutrelant, C. (2020). Deep learning-based methods for individual recognition in small birds. Methods in Ecology and Evolution, 11(9), 1072–1085. https://doi.org/10.1111/2041-210X.13436

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. Information and Software Technology, 162, 107268. https://doi.org/10.1016/j.infsof.2023.107268

Harris, C., Finn, K. R., Kieseler, M. L., Maechler, M. R., & Tse, P. U. (2023). DeepAction: a MATLAB toolbox for automated classification of animal behavior in video. Scientific Reports 2023 13:1, 13(1), 1–19. https://doi.org/10.1038/s41598-023-29574-0

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

Kholiavchenko, M., Kline, J., Ramirez, M., Stevens, S., Sheets, A., Babu, R., Banerji, N., Campolongo, E., Thompson, M., Van Tiel, N., Miliko, J., Bessa, E., Duporge, I., Berger-Wolf, T., Rubenstein, D., & Stewart, C. (2024). KABR: In-situ dataset for Kenyan animal behavior recognition from drone videos. In In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (pp. 31–40). https://doi.org/10.1109/WACVW60836.2024.00011

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

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences, 115(25), E5716-E5725. https://doi.org/10.1073/pnas.1719367115

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

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution, 14(4), 994-1016. https://doi.org/10.1111/2041-210X.14061

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

Ultralytics. (2023). YOLOv8 - Ultralytics. https://docs.ultralytics.com

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial intelligence review, 53(8), 5929-5955. https://doi.org/10.1007/s10462-020-09838-1

Wilson, O., Schoeman, D., Bradley, A., & Clemente, C. (2025). Practical guidelines for validation of supervised machine learning models in accelerometer based animal behaviour classification. Journal of Animal Ecology. https://doi.org/10.1111/1365-2656.70054

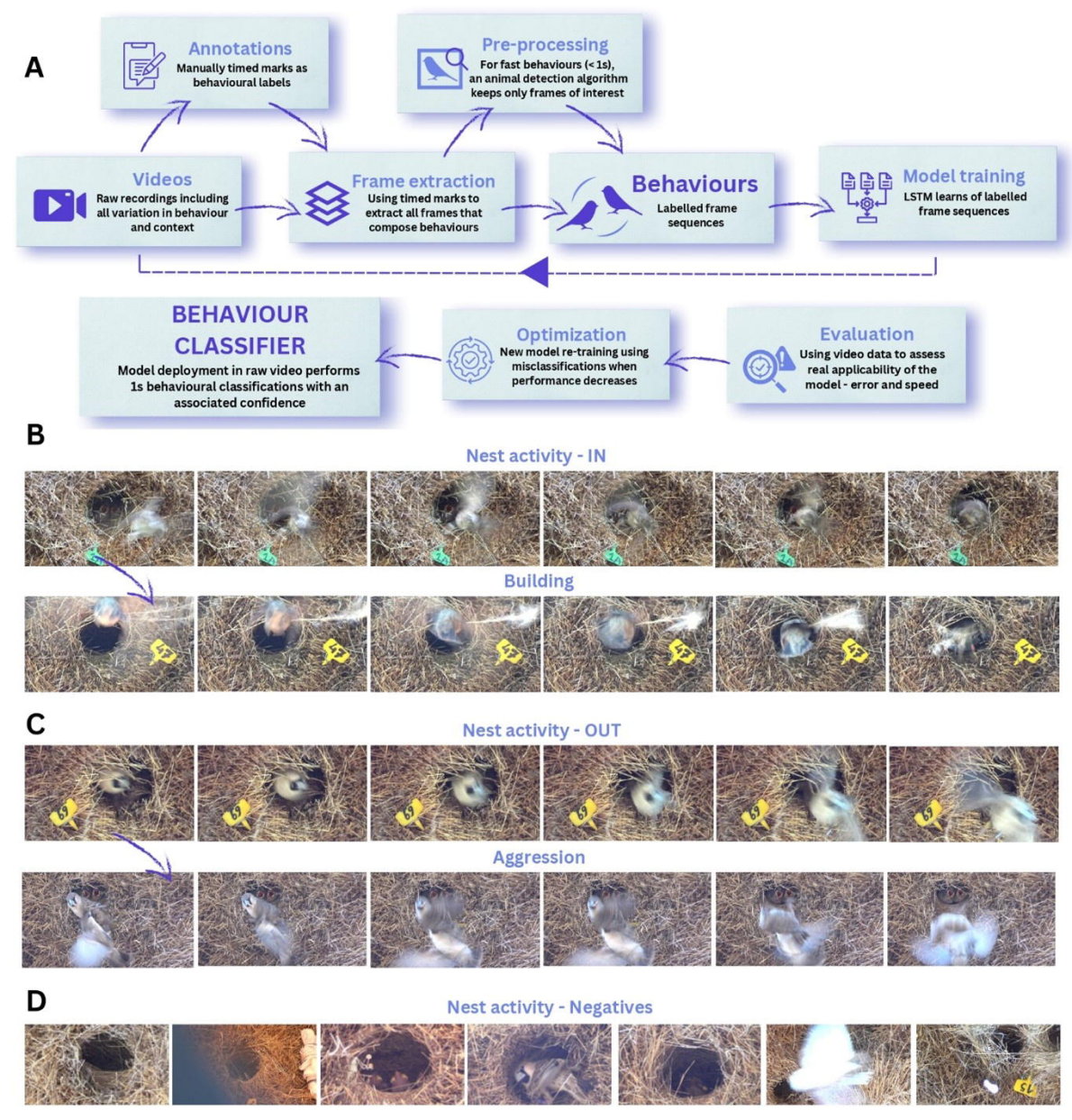
Weinstein, B. G. (2018). Scene-specific convolutional neural networks for video-based biodiversity detection. Methods in Ecology and Evolution, 9(6), 1435–1441. https://doi.org/10.1111/2041-210X.13011

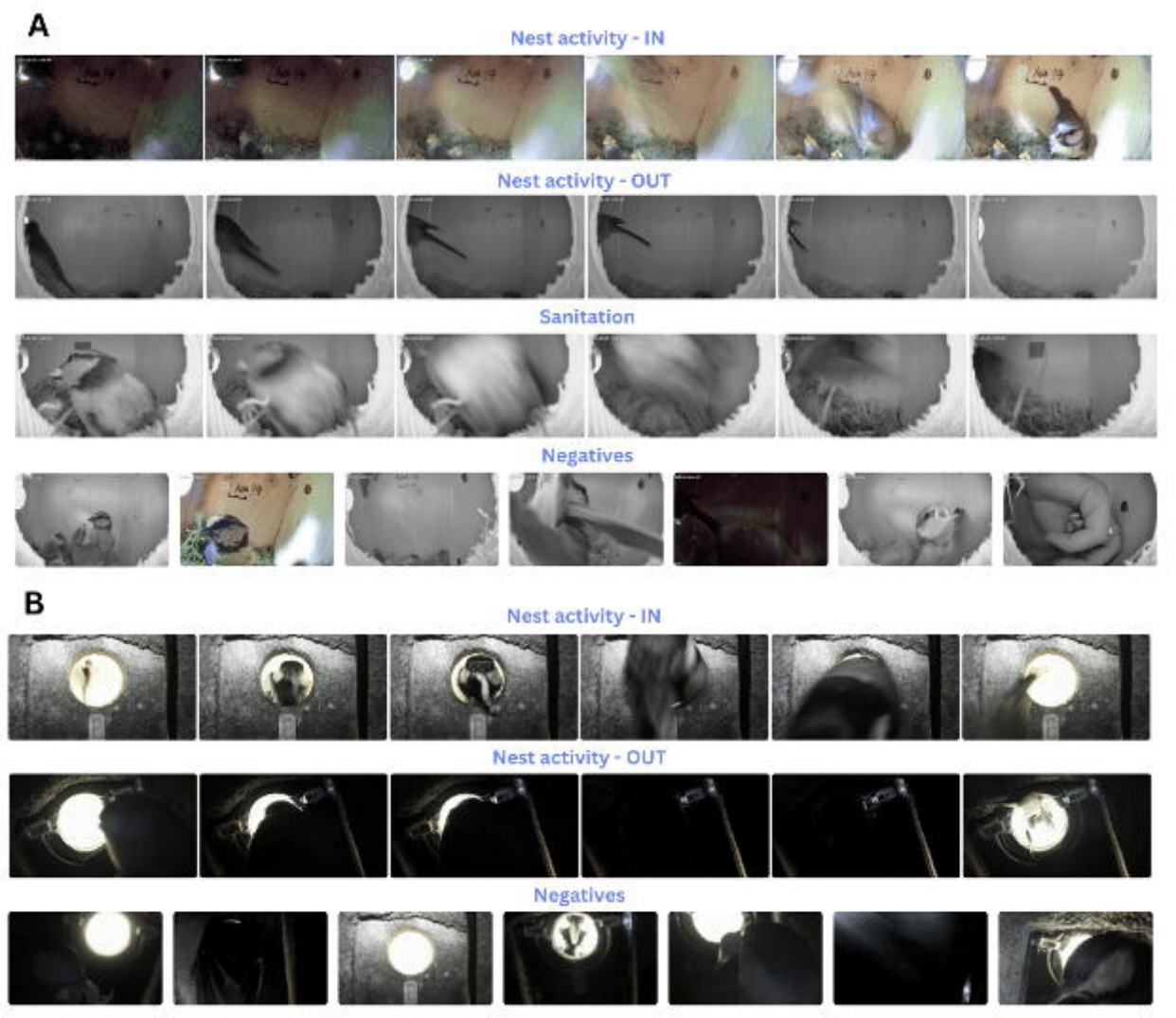
Williams, H. M., & DeLeon, R. L. (2020). Deep learning analysis of nest camera video recordings reveals temperature-sensitive incubation behavior in the purple martin (*Progne subis*). Behavioral Ecology and Sociobiology, 74(1), 1–12. https://doi.org/10.1007/S00265-019-2789-2/FIGURES/4

Table 1: Behavioural detection models' training details and performance. Training details include: classes, interval of frames, training and validation datasets sizes. Performance was assessed as i) the model frame-accuracy obtained at the validation stage, but also, ii) using videos (2h) to assess the produced false positives ("FP") and false negatives ("FN"), as well as speed (i.e., annotation time), performed by model and observers, to assess deployment success.

[a]Nest-chamber entrance, [b]nest-chamber exit, [c]negative class, *every three frames, **every six frames, □"leave-one-out" approach (SM-8D), ‡performed simultaneousy

Species	*Model*	Classes	Frames interval	Sequences training	Sequences validation	Processing	Model accuracy % (sequences)	Model FP% | FN% (videos)	Observers FP% | FN% (videos)	Model speed (videos)	Observer speed (videos)
Sociable weavers	*Nest activity*	IN[a] vs. OUT[b] vs. NC[c]	6	40,149	3,609	full video	92.02	6.24 | 0.39	10.05 | 3.36	28.6 min	40.45 min
	Building	building vs. NC[c]	6	13,122	424	only IN	87.5	3.97 | 2.3	0.58 | 4.97	< 1 min	17.73‡ min
	Aggression	aggression vs. NC[c]	6*	1,524	192	only OUT	88.02	5.82 | 0.06	1.62 | 3.22	< 1 min	
Blue tits	*Nest activity*	IN[a] vs. OUT[b] vs. NC[c]	6	9,885	2,604	full video	98.6	7.22 | 6.96	-	37.5 min	-
	Sanitation	sanitation vs. NC[c]	6	5,544	614	only OUT	92.8	7.22 | 6.74	-	< 1 min	-
Great tits	*Nest activity*	IN[a] vs. OUT[b] vs. NC[c]	6**	3,887□	3,474□	full video	93.6	5.83 | 4.73	-	26.7 min	-


A
Annotations
Manually timed marks as behavioural labels
Pre-processing
For fast behaviours (< 1s), an animal detection algorithm keeps only frames of interest
Videos
Raw recordings including all variation in behaviour and context
Frame extraction
Using timed marks to extract all frames that compose behaviours
Behaviours
Labelled frame sequences
Model training
LSTM learns of labelled frame sequences
BEHAVIOUR CLASSIFIER
Model deployment in raw video performs 1s behavioural classifications with an associated confidence
Optimization
New model re-training using misclassifications when performance decreases
Evaluation
Using video data to assess real applicability of the model - error and speed
B
Nest activity - IN
Building
C
Nest activity - OUT
Aggression
D
Nest activity - Negatives


A
Nest activity - IN
Nest activity - OUT
Sanitation
Negatives
B
Nest activity - IN
Nest activity - OUT
Negatives