Received: 24 September 2020 | Accepted: 25 January 2021

DOI: 10.1111/2041-210X.13576

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Robust ecological analysis of camera trap data labelled by a machine learning model

Robin C. Whytock[1,2] | Jędrzej Świeżewski[3] | Joeri A. Zwerts[4] | Tadeusz Bara-Słupski[4] | Aurélie Flore Koumba Pambo[2] | Marek Rogala[3] | Laila Bahaa-el-din[5] | Kelly Boekee[6,7] | Stephanie Brittain[8,9] | Anabelle W. Cardoso[10] | Philipp Henschel[11,12] | David Lehmann[1,2] | Brice Momboua[2] | Cisquet Kiebou Opepa[13] | Christopher Orbell[1,11] | Ross T. Pitman[11] | Hugh S. Robinson[11,14] | Katharine A. Abernethy[1,12]

[1]Faculty of Natural Sciences, University of Stirling, Stirling, UK; [2]Agence Nationale des Parcs Nationaux, Libreville, Gabon; [3]Appsilon AI for Good, Warsaw, Poland; [4]Utrecht University, Utrecht, The Netherlands; [5]School of Life Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa; [6]Program for the Sustainable Management of Natural Resources, South West Region, Buea, Cameroon; [7]Center for Tropical Forest Science, Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama; [8]Department of Zoology, The Interdisciplinary Centre for Conservation Science, University of Oxford, Oxford, UK; [9]The Institute of Zoology, Zoological Society of London, London, UK; [10]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA; [11]Panthera, New York, NY, USA; [12]Institut de Recherche en Ecologie Tropicale, CENAREST, Libreville, Gabon; [13]Wildlife Conservation Society, Kinshasa, Republic of the Congo and [14]Wildlife Biology Program, W.A. Franke College of Forestry and Conservation, University of Montana, Missoula, MT, USA

Correspondence
Robin C. Whytock
Email: robbie.whytock1@stir.ac.uk

Funding information
Google Cloud for Education; EU 11th FED ECOFAC6; University of Oxford Hertford College Mortimer May Fund

Handling Editor: Carlos Alberto Silva

Abstract

1. Ecological data are collected over vast geographic areas using digital sensors such as camera traps and bioacoustic recorders. Camera traps have become the standard method for surveying many terrestrial mammals and birds, but camera trap arrays often generate millions of images that are time-consuming to label. This causes significant latency between data collection and subsequent inference, which impedes conservation at a time of ecological crisis. Machine learning algorithms have been developed to improve the speed of labelling camera trap data, but it is uncertain how the outputs of these models can be used in ecological analyses without secondary validation by a human.
2. Here, we present our approach to developing, testing and applying a machine learning model to camera trap data for the purpose of achieving fully automated ecological analyses. As a case-study, we built a model to classify 26 Central African forest mammal and bird species (or groups). The model generalizes to new spatially and temporally independent data (n = 227 camera stations, n = 23,868 images), and outperforms humans in several respects (e.g. detecting 'invisible' animals). We demonstrate how ecologists can evaluate a machine learning model's precision and accuracy in an ecological context by comparing species richness, activity patterns

Robin C. Whytock and Jędrzej Świeżewski contributed equally to the manuscript.

(*n* = 4 species tested) and occupancy (*n* = 4 species tested) derived from machine learning labels with the same estimates derived from expert labels.
3. Results show that fully automated species labels can be equivalent to expert labels when calculating species richness, activity patterns (*n* = 4 species tested) and estimating occupancy (*n* = 3 of 4 species tested) in a large, completely out-of-sample test dataset. Simple thresholding using the Softmax values (i.e. excluding 'uncertain' labels) improved the model's performance when calculating activity patterns and estimating occupancy but did not improve estimates of species richness.
4. We conclude that, with adequate testing and evaluation in an ecological context, a machine learning model can generate labels for direct use in ecological analyses without the need for manual validation. We provide the user-community with a multi-platform, multi-language graphical user interface that can be used to run our model offline.

KEYWORDS
artificial intelligence, biodiversity, birds, Central Africa, mammals

1 | INTRODUCTION

The urgent need to understand how ecosystems are responding to rapid environmental change has driven a 'big data' revolution in ecology and conservation (Farley et al., 2018). High resolution ecological data are now streamed in real-time from satellites, Global Positioning System tags, bioacoustic detectors, cameras and other sensor arrays. The data generated offer considerable opportunities to ecologists, but challenges such as data processing, data storage and data sharing cause latency between data gathering and ecological inference (i.e. creating derived ecological metrics, testing ecological hypotheses and quantifying ecological change), sometimes in the order of years or more. For example, *c*. 40% of respondents from an international survey of camera trap users regarded data analysis and cataloguing images as an important or extremely important methodological barrier (Glover-Kapfer et al., 2019). Overcoming these challenges could open the gateway to ecological 'forecasting', where directional changes in ecological processes are detected in real time and near-term, future change is predicted effectively using an iterative data gathering, model updating and model prediction approach (Dietze et al., 2018).

Digital camera traps or wildlife 'trail cams' have revolutionized wildlife monitoring and are now the 'gold standard' for monitoring many medium to large terrestrial mammals (Glover-Kapfer et al., 2019). Animals and their behaviour are identified in images either by manual labelling, using citizen science platforms (Swanson et al., 2015) or, more recently, by using machine learning models (Norouzzadeh et al., 2018; Schneider et al., 2018; Tabak et al., 2019; Willi et al., 2019). Machine learning models can at minimum separate true animal detections from non-detections (Wei et al., 2020) or in more sophisticated examples identify species, count individuals and describe behaviour (Norouzzadeh et al., 2018). These advances in machine learning have increased the speed at which camera trap data are labelled and analysed but, in all cases we are aware of, the outputs (e.g. species labels) are not used to make ecological inference directly. Instead, machine learning models are typically used as a 'first pass' to identify and group images belonging to individual species for full or partial manual validation at a later stage, or to cross-validate labels from citizen science platforms (Willi et al., 2019). These fully or partially validated labels are then used for ecological analyses. Thus, although machine learning models are reducing manual labelling times, ecologists are not yet comfortable using the outputs (e.g. species labels) as part of a completely automated workflow, from labelling to analyses. This is despite the development of advanced machine learning models that classify species in camera trap images with accuracy that (with some limitations) matches or exceeds humans (Schneider et al., 2019; Tabak et al., 2019; Willi et al., 2019).

One significant challenge limiting the application of machine learning models to camera trap data is that models rarely generalize well to completely out-of-sample data (i.e. data from new, spatially and temporally independent studies), particularly when used to classify animals to species level (Beery et al., 2018; Schneider et al., 2020). Models can quickly learn the features of specific camera 'stations' (the spatial replicate in camera trap studies) such as the general background instead of learning features of the animal itself. This problem is further amplified by the fact that rare species in the training data might only ever appear at a limited number of camera stations, so training and validation data are rarely independent. Various approaches can be used to reduce these biases, such as carefully ensuring that training and validation data are independent (e.g. by using data from multiple studies), by using data augmentation such as adding noise to training data in the form of image transformations, focusing model optimization on rare species classes and by using spatial k-fold cross-validation (Schneider et al., 2020). Until the problem of generalization can be overcome, machine learning models for classifying camera trap images will remain an important tool for reducing manual labelling effort, but they will not achieve their full potential for creating fully automated pipelines for data analysis.

Machine learning models also have the potential to be deployed inside camera trap hardware in the field at the 'edge' (i.e. on microcomputers installed inside hardware that collects data), with summarized results (e.g. species labels) transmitted in real-time via a Global System for Mobile Communications networks or via satellite (Glover-Kapfer et al., 2019). In geographically remote areas or time-sensitive situations (e.g. law enforcement) this would greatly reduce the latency between data capture and interpretation, and reduce the expense and effort required to collect data in remote regions by removing the need to transfer data-heavy images across wireless networks. However, before 'smart' cameras become a reality, it is essential that users understand how uncertainty in machine learning model predictions might impact derived ecological metrics and analyses, which are often sensitive to biases (e.g. false positives in occupancy models; Royle & Link, 2006). To achieve this, there is a need to develop workflows that test the performance of machine learning models in an ecological modelling context that goes beyond simple measures of precision and accuracy.

Ideally, if machine learning models had 100% precision and accuracy (e.g. for species identification), camera trap data could be collected, labelled automatically using the model and the results used to directly calculate ecological metrics or as variables in ecological models. However, the reality is that machine learning models are imperfect (Schneider et al., 2019). It is therefore uncertain what levels of precision and accuracy are needed to meet the requirements of ecological analyses. This is particularly the case for the spatial and temporal analyses of animal distributions in camera trap data, which require specialized ecological models (e.g. occupancy models) that account for imperfect detection (MacKenzie et al., 2003).

In this paper, we describe the approach used to build a machine learning model that identifies species in camera trap images (26 species/groups of Central African forest mammals and birds) and which generalizes to spatially independent data. To evaluate how well the machine learning model labelling precision and accuracy performs in an ecological modelling context, we (a) evaluate how uncertainties in the precision and accuracy of machine learning labels affect ecological inference (derived metrics of species richness, activity patterns and occupancy) compared to the same metrics calculated using expert, manually generated labels and (b) demonstrate a workflow to 'ground truth' the performance of machine learning models for camera trap data in an ecological modelling context. We discuss the implications of these results for making fully automated ecological inference from camera trap data using the outputs of machine learning models. We also provide the user community with an easily installed, open-source graphical user interface that needs no understanding of machine learning to run the model offline on both camera trap images and videos.

2 | MATERIALS AND METHODS

2.1 | Data preparation

As a case study, the model was developed for classifying terrestrial forest mammals and birds in Central Africa (see Table S1 for further details on species and groups), where camera traps are now frequently deployed over large spatial scales to survey secretive birds and mammals in remote and inaccessible landscapes (Bahaa-el-din & Cusack, 2018; Bessone et al., 2020; O'Brien et al., 2020). Training data were obtained from multiple countries and sources (*c*. 1.6 million images; reduced to $n = 347{,}120$ images after data processing; Table 1). Each source used different camera trap models (Reconyx, Bushnell, Cuddeback, Panthera Cams) and images were diverse in resolution, quality (e.g. sharpness, illumination) and colour. Individual studies also used different field protocols for camera deployment but all were focused on detecting terrestrial forest mammals, with cameras installed on trees approximately 30–40 cm above ground level. The exception to this was data from (Cardoso et al., 2020) who installed cameras at a height of approximately 1 m for the purpose of detecting forest elephants *Loxodonta cyclotis*. Camera trap configuration was set to be highly sensitive in some cases and images were often captured in a series of rapid, short bursts (e.g. taking 10 images consecutively). This resulted in long sequences of very similar images, for example showing an animal walking in front of the camera (Figure S1).

It was important to account for image sequences when selecting a validation set during the model training phase, since there was a risk of highly similar images being present in both the training and validation sets. To address this issue, the training and validation split was performed based on image metadata (timing of images and image source) to identify unique 'events' and camera locations that were

TABLE 1 Sources of training data used to train the machine learning model for classifying species in camera trap images, sorted by number of images provided. The final subset of data used to train the model was $n = 347{,}120$ images (see later)

Source	Country	Reference	*n* images	*n* unique camera locations
Anabelle Cardoso	Gabon	Cardoso et al. (2020)	102,418	40
Kelly Boekee	Cameroon	—	123,954	60
Cisquet Kiebou Opepa	Republic of Congo	—	60,393	64
Joeri Zwerts	Cameroon	—	36,027	30
Laila Bahaa-el-Din	Gabon	Bahaa-el-din et al. (2013)	16,558	40
Stephanie Brittain	Cameroon	—	7,770	32

not replicated across the training and validation split (Norouzzadeh et al., 2019). This approach posed a challenge for maintaining class balances in the training and validation sets, but it reduced the risk of non-independent training and validation sets. A total of 27 classes were used to train the model, which were mostly mammals or mammal groups ($n = 21$), birds ($n = 4$), humans ($n = 1$) and 'blank' images (i.e. no mammal, bird or human). Details of taxonomy and justification for species groups are in Table S1.

2.2 | Issues identified in the training data

Our 'real-life' training data had not been pre-processed or professionally curated for the purposes of training machine learning models and naturally contained errors that arise from hardware faults, human error and different approaches to manual species labelling by experts. We identified three primary sources of error. The first was over-exposed images (a hardware fault) where the image foreground was 'flooded' by the flash (usually at night), making the image appear mostly white. Animals in these images were sometimes partially visible and could be classified by a skilled human observer, despite the loss of colour information, texture and other detail. However, over-exposed images presented a challenge for the machine learning model because white dominated the image regardless of the species.

The second main source of error was caused by under-exposed images. This error was revealed after inspecting model outputs during the training phase, and showed that highly under-exposed images appeared almost entirely or entirely black to a human observer, but the machine learning model was capable of using information in the image to detect and correctly classify the species (Figure 1).

The final source of error in the training data was mislabelled images (e.g. confusing similar species, such as chimpanzee *Pan troglodytes* and gorilla *Gorilla gorilla*) and using different approaches to labelling, for example one data source combined all primates into 'monkey', whereas other data sources separated apes from other primates.

We used an iterative approach to address these issues that consisted of model training, validation, error correction (correcting mislabelled images in the training data) and model updating. In particular, we carefully inspected images that appeared to be incorrectly labelled by the model, but which were labelled with high confidence. This approach revealed hidden problems in the data, such as the presence of animals in under-exposed images that would have otherwise led us to underestimate the model's performance.

2.3 | Machine learning model

Our primary objectives were to demonstrate and test how the outputs of a machine learning model can be evaluated in an ecological context and used directly in ecological analyses without manual validation. We therefore summarize our approach to building the machine learning model here. Full details of the machine learning training scheme and implementation can be found in Supporting Information and at our GitHub repository (Świeżewski & Whytock, 2021, 2021).

In summary, we used the established ResNet50 architecture to build the model (He et al., 2016) and transfer learning was used to speed up training with weights pre-trained on the ImageNet dataset. We identified species using the entire image frame without using bounding boxes and used basic augmentation (horizontal flips, rotations, zoom, lighting and contrast adjustments, and warps) during training, but not during model validation. We used one-cycle policy training (Smith, 2018) and trained using progressive resizing in two stages. These approaches were implemented using the Fastai Python library (Howard & Gugger, 2020).

2.4 | Out-of-sample test data

One of the major limitations to model performance for camera trap images is the ability to generalize predictions to new, independent camera stations, that is, unique locations with different backgrounds not seen during model training (Beery et al., 2018; Schneider et al., 2020). Since we aimed to create a model that could generalize well to new camera locations, we tested the final model's performance using a new out-of-sample dataset that was completely spatially and temporally independent from the data used to

FIGURE 1 (Left) Raw image from the dataset, labelled by experts as 'blank', but classified by the machine learning model with high certainty as a red duiker. (Right) The same image but manually brightened by narrowing the displayed colour spectrum, reveals a red duiker is present and the model was correct

train and validate the model. These out-of-sample data consisted of 23,868 images from 227 unique camera locations surveyed between 16 January 2018 and 4 October 2019 in Gabon from three distinct study areas totalling 3,701 km^2 of forest (Orbell & Whytock, 2021). Cameras also differed from the models used in the training data (Panthera Cams V4 and V5), but field protocols were similar and cameras were placed approximately 30 cm above the ground on a tree at a distance of *c*. 3–5 m perpendicular to the centre of animal trails. Single-frame images were captured using medium sensitivity settings, and images were separated by a minimum of 1 s. The aim of the study was to survey the small-to-large mammal community, with a particular focus on great apes (*Pan troglodytes*, *Gorilla gorilla*), forest elephants *Loxodonta cyclotis*, leopard *Panthera pardus* and African golden cat *Caracal aurata*. These data (n = 227 camera stations, n = 23,868 images, median 75, range 1–545 images per station) were manually labelled by an expert (co-author CO).

2.5 | Summary of model's general performance

To allow general comparison of our model's performance with other similar models in the literature (Norouzzadeh et al., 2018; Schneider et al., 2018; Tabak et al., 2019; Willi et al., 2019) we calculated top one and top five accuracies using the out-of-sample data (n = 227 camera stations). Top one accuracy is the percent of expert labels that match the top-ranking label generated by the machine learning model. Top five accuracy calculates the percent of expert labels that match any of the top five ranking machine learning generated labels. Top one accuracy for the overall machine learning model was 77.63% and top five accuracy was 94.24% (Table S2; Figures S2, S3 and S4). After aggregating labels of similar species that were frequently mis-classified by the model into a reduced set of 11 classes, top one and top five accuracies increased to 79.92% and 95.99%, respectively (Figures S5 and S6). The model can classify around 4,000 images (*c*. 0.5 MB in size) per hour using an Intel® Core™ i7-8665U CPU @ 1.90 GHz × 8. For comparison, based on our experience, manual labelling can be done at speeds ranging from 125 to 500 images per hour depending on the quality of the images and if images are captured in sequences (which can be faster to label manually).

We also compared the precision and recall for each species from our optimal model with the precision and recall for the same species reported for the model used by the WildlifeInsights web-platform (www.wildlifeinsights.org; Ahumada et al., 2020). This global project uses a deep convolutional neural network trained using Google's Tensorflow framework and a training dataset of 8.7 M images, comprising 614 species.

2.6 | Comparing derived ecological metrics using machine learning labels and expert labels

We calculated three common ecological metrics for the out-of-sample data (raw species richness at individual camera stations, activity patterns for four focal species, and occupancy for four focal species) separately using the manually generated, expert labels and the machine learning generated labels. Species richness (the number of species in a discrete unit of space and time) can be used to quantify temporal and spatial changes in biodiversity. Although other measures of species diversity exist, we chose this simple metric because it is widely used in the ecology literature despite its limitations. Activity patterns describe the diel activity patterns of focal species (Rowcliffe et al., 2014) and are typically calculated using camera trap data to understand fundamental life history traits and behaviour such as temporal niche partitioning. Occupancy models are hierarchical models commonly fitted to camera trap data because they can account for imperfect detection (which rarely equals 1) to estimate the conditional probability that a site is 'occupied' by a species given it was not detected (MacKenzie et al., 2002, 2003). Covariates such as measures of vegetation cover can be included in both the detection and occupancy component models. These models are relatively complex, and small changes in detection histories (presence or absence of a species during a discrete time interval), false positives or false negatives can dramatically affect results (Royle & Link, 2006). We therefore predicted that occupancy estimates obtained using machine learning generated labels would compare poorly with estimates using expert, manually generated labels. Other commonly used metrics such as spatially explicit capture recapture methods (Borchers & Efford, 2008) and the random encounter model (Lucas et al., 2015) were not evaluated because they either required an additional layer of analysis (e.g. individual identification and re-identification), or because the sampling design of our test data was unsuitable (e.g. non-random camera placement).

The four focal species used for calculating activity patterns and occupancy were African golden cat, chimpanzee, leopard and African forest elephant. These species were chosen because they were the focus of the camera trap survey that generated the out-of-sample test data and because they are conservation priority species in Central Africa. We also initially included western lowland gorilla, but we had too few unique captures of this species (only seven of 227 out-of-sample stations having >5 captures) to fit either activity pattern models or occupancy models.

2.7 | Thresholding and overall model performance

All three metrics derived from machine learning labels were re-calculated using a threshold approach, where labels were excluded if the model's predicted 'confidence' (softmax output) was below a given threshold. While these softmax values are not strictly describing the model's 'confidence', and they are fallible under malicious attack (Guo et al., 2017; Kurakin et al., 2017), they do in general correlate with prediction accuracy (e.g. see Results). For the sake of brevity we refer to these values as 'confidence' hereafter. The thresholds tested ranged from 0 (no threshold) to 90%, increasing in 10% intervals. For each of the three ecological metrics, we then

re-calculated results using the machine learning labels and compared these with results from the full, expert labelled dataset using various statistical measures (see later). Secondarily, we also made the same comparison using the matching subset of machine learning labels and expert labels after thresholding, but this was considered less challenging for the model. We also calculated the effect of removing data on sample size, top one balanced accuracy and top five accuracy for the overall model, and on four standard measures of model precision and accuracy (precision, recall, F1 score and balanced accuracy for each species using the confusionMatrix function in the CARET R package (Kuhn, 2020).

Estimated species richness from machine learning generated labels and expert labels was compared using linear regression fitted by least squares. Species richness from expert labels was used as the predictor variable and species richness from machine learning labels was used as the response. For each threshold, we evaluated how well species richness from machine learning labels correlated with expert labels by calculating the slope coefficient and variance explained (R^2).

Diel activity patterns (proportion of 24 hr day active) were calculated using circular kernel probability density functions for all four focal species, fitted using the fitact function (with 200 bootstrap replicates) from the ACTIVITY R package (Rowcliffe, 2019; Rowcliffe et al., 2014). For each species and threshold combination, we tested if there was a significant difference (Wald test on chi-squared distribution with 1 degree of freedom) in diel activity estimated by machine learning labels and expert labels using the compareAct function (Rowcliffe, 2019), expecting no difference using an alpha level of 0.05.

Single season, single species occupancy models were fitted by maximum likelihood using the occu function from the UNMARKED R package (Fiske & Chandler, 2011), where the occupancy state Z_i (1 = occupied, 0 = unoccupied) of a site M (camera station) is modelled as:

$$Z_i \sim \text{Bernoulli}(\psi) \text{ for } i = 1, 2, \ldots, M,$$

and the observation process Y on sampling occasion i at site j as:

$$Y_{ij} \mid Z_i \sim \text{Bernoulli}(Z_i p) \text{ for } j = 1, 2, \ldots, J_i.$$

Detection histories were collapsed to 5-day sampling occasions as a compromise between achieving model stability and ensuring an adequate number of replicates for each site. In the detection (observation process) model, we included covariates (using a logit link) of Elevation (m), Date (first day of the 5-day occasion length) and Date^2 (to allow for non-linear, seasonal changes in detection). In the occupancy component model, we included covariates of Elevation (m), Distance to the Nearest River (m), Distance to the Nearest Road (m) and mean distance to the Nearest Village (m) as continuous predictors without interactions. All covariates were mean-centred and scaled by 1 *SD* to prevent convergence issues. We did not perform model selection and predicted occupancy for the 227 camera stations using the full model.

We then compared occupancy predictions (n = 227 camera stations) for no threshold (i.e. using all data), and the nine thresholds using linear regression fitted by least squares as described previously for the species richness comparisons.

3 | RESULTS

3.1 | Effect of thresholding on overall model performance

Regardless of the threshold used, top five accuracy for the overall model predictions on the out-of-sample data (n = 227 independent camera stations) were consistently close to or above 95% (Figure 2). To achieve a top one balanced accuracy of 90% or more for the overall model, a threshold of ≥70% confidence was required and >25% of the data were discarded (Figure 2).

Table 2 shows performance statistics for each class compared to the WildlifeInsights model. Note that this comparison should be interpreted with caution. Ideally, we would run the WildlifeInsights model on our out-of-sample test data, but data sharing restrictions prevented this. Where our species or groups could not be compared with an equivalent class on WildlifeInsights this is indicated as no equivalent class (NE). If precision and recall could not be estimated because of insufficient training and validation data this is indicated as 'needs more data' (NMD). With a threshold of 70% confidence (i.e. excluding labelled images below 70% confidence), top one balanced accuracies for 16 of the 27 classes were >90% and a further five were >75% (Table 2). Top one balanced accuracies for the remaining seven classes ranged

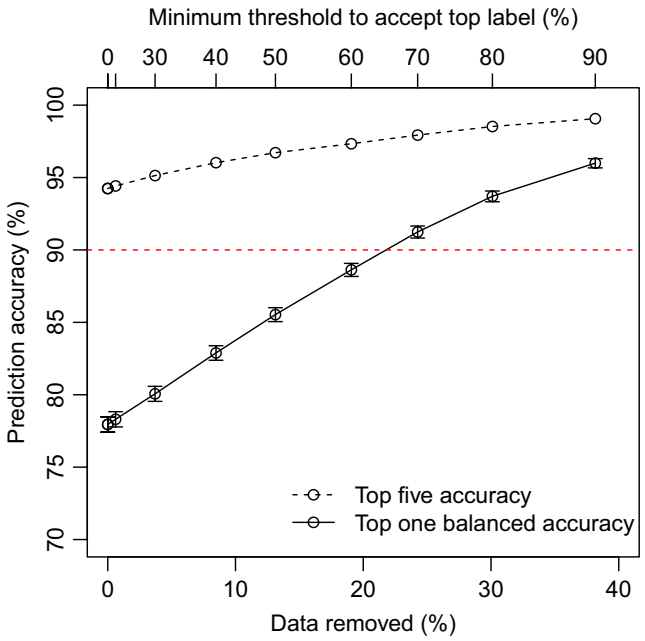

FIGURE 2 Relationship between threshold level to accept top label, % of data discarded and overall top five and top one balanced accuracy (±95% CI) for predictions on out-of-sample test data

Species	Precision %	Recall %	F1	Prevalence	Balanced accuracy
Civet_African_Palm	NMD (NMD)	NMD (NMD)	NA	NA	NA
Gorilla	NMD (NMD)	NMD (NMD)	NA	0.4	50
Rail_Nkulengu	0.0 (47.2)	0.0 (48.6)	NA	NA	50
Guineafowl_Crested[a]	100 (99.8)	5.3 (91.2)	10	0.1	52.6
Mandrillus	83.9 (96.1)	29 (72.3)	43.1	1.8	64.5
Blank	98.1 (98.3)	40.3 (78.7)	57.1	3.6	70.1
Buffalo_African	97.5 (91.1)	55.7 (73.6)	70.9	1.2	77.8
Bird	11.2 (NE)	60.0 (NE)	18.9	0.1	79.7
Chevrotain_Water	100 (NMD)	67.4 (NMD)	80.6	0.2	83.7
Guineafowl_Black	70.6 (79.6)	72.7 (79.5)	71.6	0.2	86.3
Cat_Golden	96.0 (NMD)	78.0 (NMD)	86.1	1	89
Pangolin	94.1 (NMD)	80.0 (NMD)	86.5	0.1	90
Duiker_Yellow_Backed	97.5 (88.8)	83.8 (72.3)	90.2	2.9	91.9
Human	78.4 (84.8)	87.4 (75.2)	82.6	4	93.2
Chimpanzee	83.5 (87)	88.4 (71.4)	85.9	2.2	94
Monkey	70.7 (NE)	92.0 (NE)	80	2.9	95.4
Mongoose	83.5 (NMD)	91.0 (NMD)	87.1	0.4	95.5
Rat_Giant	68.2 (76)	93.8 (75.8)	78.9	0.1	96.9
Duiker_Red[b]	95.9 (95.6)	96.5 (79.6)	96.2	30.8	97.3
Duiker_Blue	90.04 (98.2)	97.0 (65.7)	93.6	17.6	97.4
Hog_Red_River	97.0 (82.7)	95.7 (84.7)	96.3	6.5	97.7
Squirrel	85.9 (98.6)	95.8 (67.6)	90.6	0.9	97.8
Leopard_African	92.8 (85.2)	96.0 (61.4)	94.4	2.2	97.9
Elephant_African	91.9 (94.4)	98.4 (84.2)	95.1	19.3	98.2
Porcupine_Brush_Tailed	93.9 (89.4)	98.9 (42.1)	96.3	0.5	99.4
Genet	95.3 (89.2)	99.3 (65.6)	97.2	0.8	99.6
Mongoose_Black_Footed	92.9 (NMD)	100 (NMD)	96.3	0.1	100

TABLE 2 Precision, recall, balanced accuracy ((sensitivity + specificity)/2), F1 score and prevalence (%s of each class) for the 27 classes (Table S1) in the out-of-sample test data after removing labels with a predicted confidence <70%. The same measures for the full dataset without any thresholding are given in Table S2. Species are sorted from lowest to highest balanced accuracy. For comparison, the precision and recall for the model used by the wildlifeinsights.org web platform are given in brackets. Orange indicates our model performed worse than the WildlifeInsights model for a given species, and purple indicates our model performed better

[a]Used precision and recall for similar *Guttera plumifera* from WildlifeInsights.

[b]Used precision and recall for *Cephalophus callipygus* from WildlifeInsights.

from 50% to 70% (Table 2). All other measures of accuracy and precision at all thresholds are in Table S3 and Figure 3 shows the confusion matrix for the out-of-sample data after excluding labels below 70% confidence (see Figure S5 for the confusion matrix of aggregated labels after thresholding).

3.2 | Species richness

Species richness estimated by machine learning labels and expert labels was strongly correlated at all thresholds used (Figure 4). There was a general tendency for species richness to be underestimated by machine learning as the threshold increased, and the slope of the relationship was close to one with no threshold. A repeat analysis comparing machine learning labels and matching subsets of expert labels for each threshold is shown in Figure S8.

3.3 | Activity patterns

Above a threshold of 70% there was no significant difference between diel activity patterns estimated by machine learning labels and expert labels for all four focal species in the full out-of-sample test data (Figure 5; Table S4). The difference became non-significant at a threshold of 50% when comparing matching subsets of machine learning labelled data and manually labelled data at each threshold (Table S5).

3.4 | Occupancy models

As expected, occupancy estimates made using machine learning labels were sometimes inconsistent with those made using expert labels, and thresholding had a strongly improved inference in some cases (Figure 6). For golden cat and leopard, which are predicted

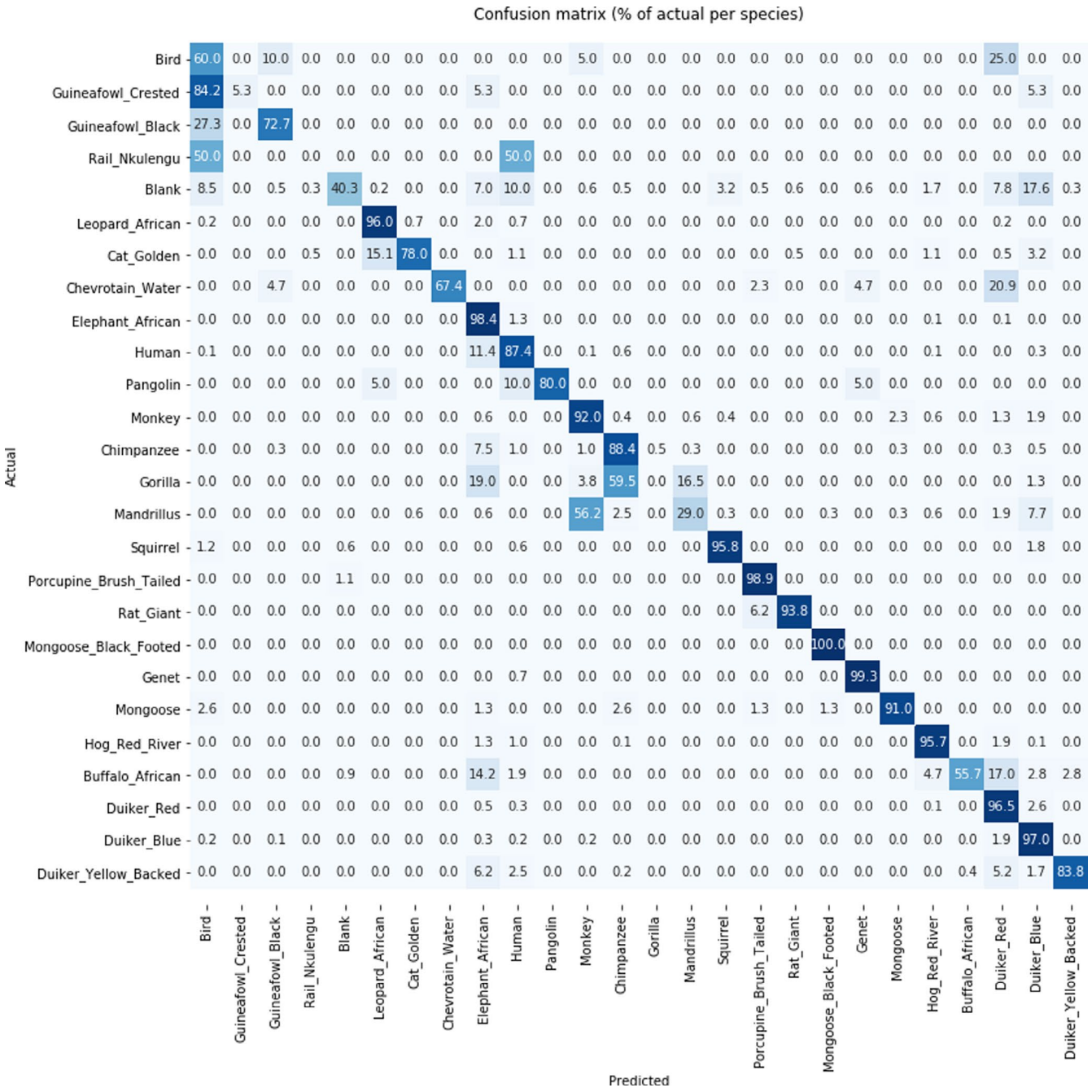

FIGURE 3 Confusion matrix (% correct labels for each species/group) showing model performance on out of sample test data after excluding labels below a confidence threshold of 70% (each row is normalized independently). Figure S7 shows the confusion matrix with absolute numbers

with high accuracy and precision by our machine learning model, occupancy estimates from machine learning labels and expert labels were highly correlated at all thresholds (Figure S9). African elephant occupancy estimates using machine learning labels improved dramatically as the threshold increased, but chimpanzee occupancy estimates from machine learning labels were consistently uncorrelated with those estimated using expert labels (Figure 6). Comparisons between matching subsets of machine learning labelled data and manually labelled data at each threshold showed much stronger correlations for all groups and thresholds (Figure S10).

4 | DISCUSSION

Machine learning models have the potential to fully automate labelling of camera trap images without the need for manual validation. This would allow ecologists to rapidly process data and use the outputs (e.g. species labels) directly in ecological analyses, but it has been uncertain how this can be achieved. In particular, models published to date do not evaluate their predictive performance in an ecological modelling context (Beery et al., 2018; Norouzzadeh et al., 2018; Tabak et al., 2019; Willi et al., 2019). Here, we compared

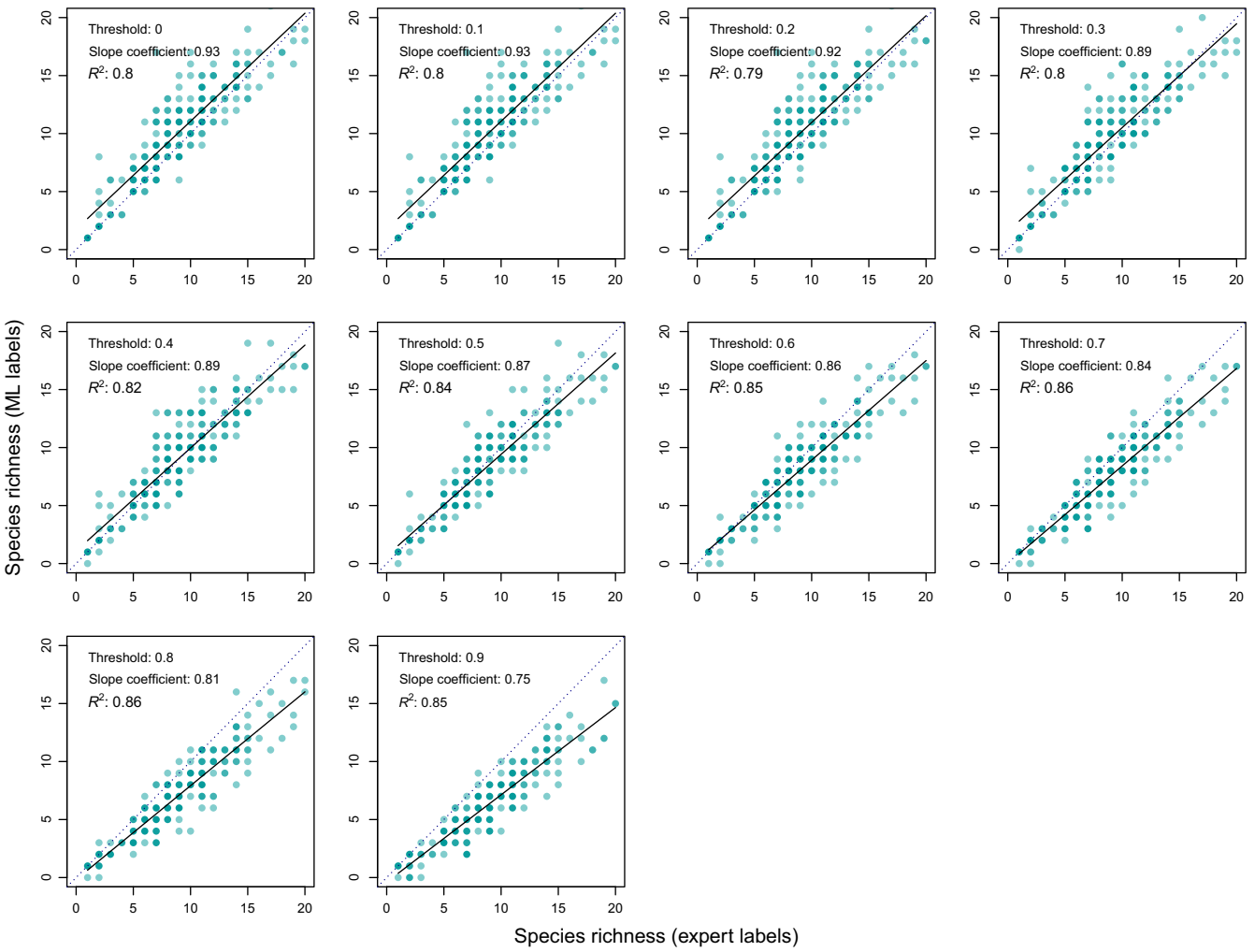

FIGURE 4 Relationship between species richness at each camera station ($n = 227$) predicted by the machine learning model (*y*-axis) and species richness predicted from expert labels (*x*-axis) for no threshold and the nine thresholds used after predicting on the out-of-sample test data (see Figure S8 for comparison between estimates from machine learning labels and matching subsets of manual data at each threshold). The dotted line shows where a 1:1 relationship would fit the data

ecological metrics calculated on an out-of-sample test dataset using machine learning labels with the same metrics calculated using expert, manually generated labels. Using our optimal species classification model that generalizes to out-of-sample data, we show machine learning labels have the potential to be used in a fully automated workflow that could remove the need for manual validation prior to conducting ecological analyses.

Most of the training approaches and many of the mechanisms we used to enhance the training of the machine learning model were taken directly or almost directly from the open source fast.ai Python library (Howard & Gugger, 2020), demonstrating the strength of this flexible library for species classification tasks. We used an established architecture (ResNet50) for the machine learning model but other more recent architectures could yield further increases in performance (Schneider et al., 2020). The ResNeXt (Xie et al., 2017), the ResNeSt (Zhang et al., 2020) and the EfficientNet (Tan & Le, 2020) families of network architectures are particularly worth exploring in this context. Another avenue of possible further improvement is to use an approach based on a sequence of models. One natural step is to first detect a bounding box for an animal with a localization model (Beery et al., 2019) and later classify only the content found in that box. Independently, another step can be introduced where a model is trained to first identify an aggregated species class (comprised of species that share similar characteristics; e.g. see Figure S6), and later dedicated models are trained to identify the individual species within these aggregated classes. However, in at least one example used to classify invertebrates in images this approach resulted in lower performance (Ärje et al., 2019).

We used a relatively small training set (*c*. 300,000 images here vs. 3.2 million in (Norouzzadeh et al., 2018) and 8.7 M used by WildlifeInsights (Ahumada et al., 2020) and a large number of individual classes, yet our model achieved relatively high precision and accuracy even when tested on completely out-of-sample data, which is considered a significant challenge for the field (Beery et al., ,2018, 2019; Schneider et al., 2019). We believe this encouraging result can be explained both by the machine learning approaches used (e.g.

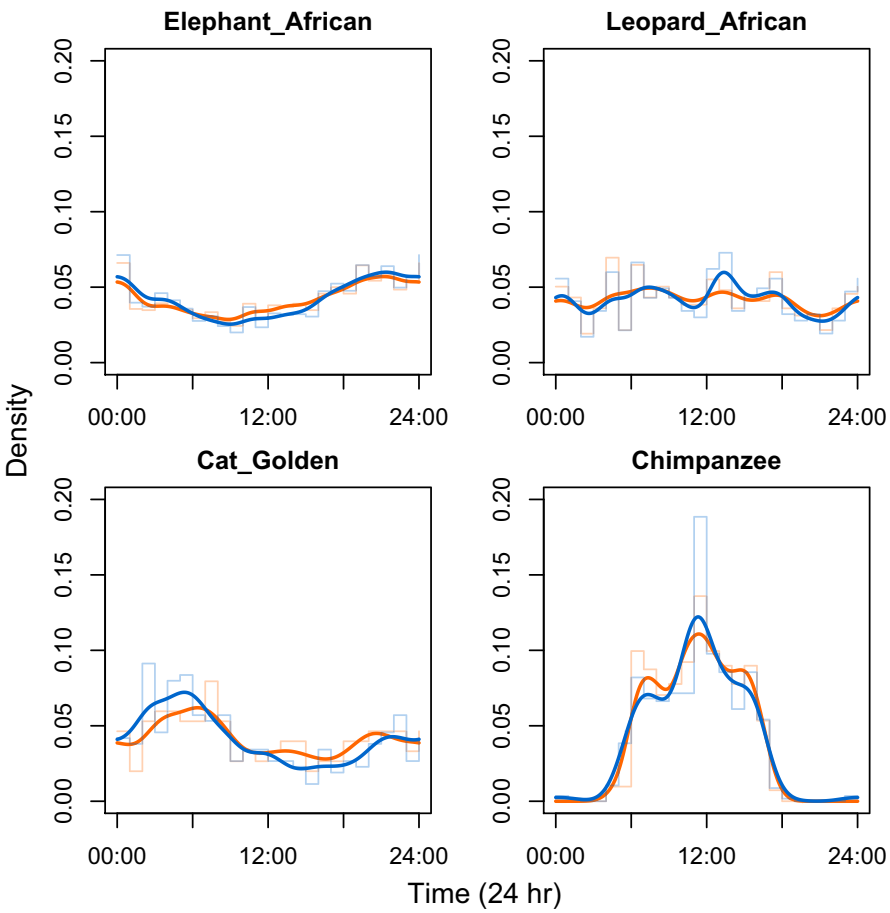

FIGURE 5 Estimated activity patterns for the four focal species in the out-of-sample test data using machine learning labels (orange; $n = 18{,}078$ observations after excluding labels below 70% confidence) and expert labels from the full dataset (blue; $n = 23{,}868$ observations)

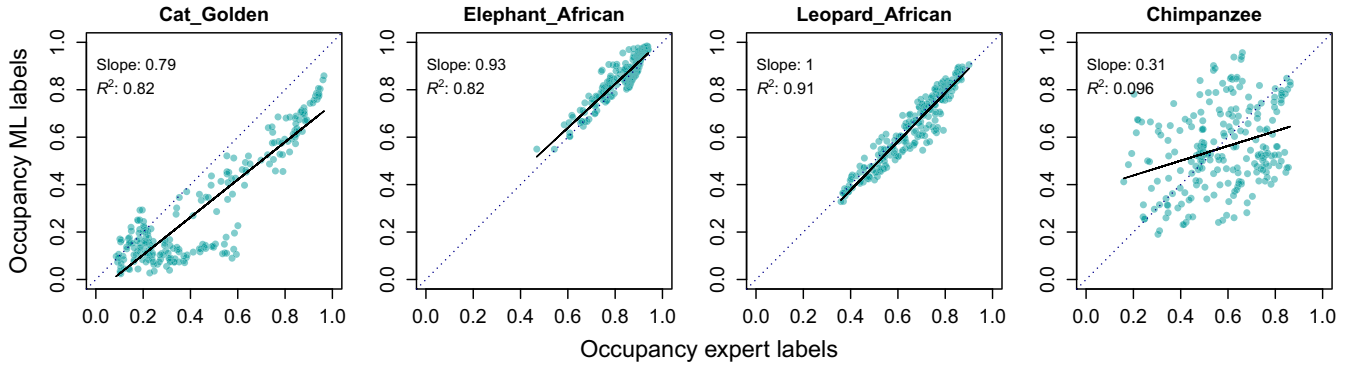

FIGURE 6 Relationship between estimated occupancy probability for $n = 227$ camera stations (points) from machine learning (ML) labels (y-axis) and expert labels (x-axis) for the four focal species after discarding labels below a 90% threshold of predicted confidence. Plots for all thresholds tested are shown in Figure S9

the fast.ai framework and image augmentation), and because forest camera traps in Central Africa are often deployed in very similar settings and using standard protocols (e.g. typically attached to trees 30–40 cm above ground level), with animals captured at a predictable distance from the camera (usually on a path) with a general background of green and brown vegetation. This is in contrast to camera trap images from more open habitats, where animals are often detected across a wide range of distances and backgrounds (Beery et al., 2018). On the other hand, informational richness in the background of photos taken in forest settings poses a significant challenge to machine learning models as well as human experts (Figure 7).

Thresholding improved the overall performance of the model and its performance for individual species. In our tests we 'discarded' labels with low 'confidence' but these data could equally be classified manually if sample sizes were small. It is important to note, however, that this additional effort to manually label low confidence images would not have improved inference in our example ecological analyses, with the exception of chimpanzee occupancy estimates. Chimpanzee images had the lowest measure of precision

FIGURE 7 An image correctly classified as nkulengu rail by our machine learning model but marked as blank by an expert. The bird is visible slightly right of centre. The dark beak is pointing left and most of the body is hidden behind branches and leaves. A section of its characteristic red legs is visible between the leaves. The model used features from the beak and head region to identify the bird (see Figure S11)

among the four focal species, which suggests that true detection events were probably missed frequently, resulting in false negatives (Figure S2). Species that were classified with the highest precision and accuracy were either relatively unique in their shape, colour and pattern (e.g. African leopard, the 'Genet' group) or were well represented in the training data. To allow users in Central Africa to use our model, we have created an offline, multi-platform software tool that can label large batches of images or videos, and display simple maps of species presence/absence and species richness (see details in Świeżewski & Whytock et al., 2021). The software also outputs the labels in a format that can be used for calculating activity patterns or for use in occupancy models. We do not fully automate these analyses at present, but we anticipate these features will be integrated into future releases.

If machine learning models can fully automate labelling of camera trap images, the first question likely to be posed by most ecologists is 'Should we?'. Camera trap images contain a wealth of information beyond species identity that would be missed using our model such as behaviour, demography, individual phenotype and body condition. A trained model is also limited to detecting and classifying the species in the training dataset, and by definition cannot detect new species. Some machine learning models can already classify behaviour (Norouzzadeh et al., 2018) and other future models will achieve this and much more. In our opinion, fully automated labels can and should be used in ecological analyses, but only after validation (and re-validation) of the model's performance, and to answer clearly defined questions. Each use-case will also differ in the benefits that can be gained from fully automated analysis. A conservation manager with tens of thousands of images collected on a rolling basis might accept a trade-off between increased speed of data analysis and having to discard images with uncertain labels, but a scientist testing hypotheses for peer-reviewed publication might prefer to view all of the images manually. We recommend that in all cases model performance should be validated regularly using sub-sampled data to detect potentially new or hidden biases. Model accuracy could change if field protocols or environmental conditions change in seasonal or unexpected ways (e.g. heavy snowfall in temperate zones). However, during model evaluation we found that expert labels in the training and validation data were also never 'perfect', and perhaps high performance machine learning models offer a more consistent means of analysing camera trap data than manual labelling because biases are predictable and can be quantified explicitly.

Camera traps are commonly used worldwide by conservation practitioners whose normal scope of work might not allow sufficient time for the handling, processing and analysing of large quantities of digital data. The authors personally know of several large camera trap databases that have not been analysed years after data collection ended, often because of a lack of resources or technical expertise. New web-based platforms for ecological data are seeking to address this problem by allowing users to upload data to the cloud where it is stored and analysed using machine learning models (Ahumada et al., 2020; Aide et al., 2013) but a lack of fast internet access can be a barrier to using such platforms and our offline application can fill this important gap. The next generation of camera traps will also have embedded machine learning models. Together, edge and cloud computing will open the door to national and international real-time ecological forecasting at unprecedented spatial and temporal scales. We anticipate that the workflow presented here could substantially improve how camera trap data are processed and analysed, and conclude that, with careful testing and evaluation in an ecological context, high performance machine learning models can be used for fully automated labelling of camera trap images.

ACKNOWLEDGEMENTS

R.C.W. was funded by the EU 11th FED ECOFAC6 program grant to the National Parks Agency of Gabon. Appsilon Data Science funded the machine learning model and software development costs. Cloud computing costs were funded by a Google Cloud Education Grant awarded to K.A.A. Camera trap data from co-authors K.B. and C.K.O. were kindly made available by the Tropical Ecology Assessment and Monitoring Network (now https://wildlifeinsights.org). Camera trap data from A.W.C. were funded by the Hertford College Mortimer May Fund at Oxford University.

AUTHORS' CONTRIBUTIONS

R.C.W., J.S., M.R., A.F.K.P., P.H., C.O., R.T.P., H.S.R., K.A.A., T.B.-S. designed research; R.C.W., J.S., M.R. performed research; R.C.W., J.S. analysed data; R.C.W., J.A.Z., L.B., K.B., A.W.C., D.L., B.M., C.K.O., C.O. collected data; R.C.W., J.S., J.A.Z., T.B.-S., A.F.K.P., M.R., L.B., K.B., S.B., A.W.C., P.H., D.L., C.O., H.S.R., K.A.A. wrote the paper.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13576.

DATA AVAILABILITY STATEMENT

All derived data used in the analyses are available online at the University of Stirling's Online Repository for Research Data http://hdl.handle.net/11667/170 (Orbell & Whytock, 2021) and the accompanying R code is available at http://doi.org/10.5281/zenodo.4486332 (Whytock, 2021). Code for the machine learning model is available online at http://doi.org/10.5281/zenodo.4534108 (Świeżewski & Whytock, 2021, 2021). Raw camera trap images used for model training and testing are from multiple individuals, institutions and sources (see Table 1 in Section 2) and are available on a case-by-case basis by request to the contributing authors.

ORCID

Robin C. Whytock https://orcid.org/0000-0002-0127-6071
Jędrzej Świeżewski https://orcid.org/0000-0001-7005-8003
Joeri A. Zwerts https://orcid.org/0000-0003-3841-6389
Marek Rogala https://orcid.org/0000-0002-9949-4551
Kelly Boekee https://orcid.org/0000-0001-8131-5204
Stephanie Brittain https://orcid.org/0000-0002-7865-0391
Anabelle W. Cardoso https://orcid.org/0000-0002-4327-7259
David Lehmann https://orcid.org/0000-0002-4529-8117
Ross T. Pitman https://orcid.org/0000-0002-3574-0063
Hugh S. Robinson https://orcid.org/0000-0002-4060-3143
Katharine A. Abernethy https://orcid.org/0000-0002-0393-9342

REFERENCES

Ahumada, J. A., Fegraus, E., Birch, T., Flores, N., Kays, R., O'Brien, T. G., Palmer, J., Schuttler, S., Zhao, J. Y., Jetz, W., Kinnaird, M., Kulkarni, S., Lyet, A., Thau, D., Duong, M., Oliver, R., & Dancer, A. (2020). Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, *47*(1), 1–6. https://doi.org/10.1017/S0376892919000298

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, *1*, e103. https://doi.org/10.7717/peerj.103

Ärje, J., Raitoharju, J., Iosifidis, A., Tirronen, V., Meissner, K., Gabbouj, M., Kiranyaz, S., & Kärkkäinen, S. (2019). Human experts vs. machines in taxa recognition. *ArXiv:1708.06899 [Cs, q-Bio, Stat]*. Retrieved from http://arxiv.org/abs/1708.06899

Bahaa-el-din, L., & Cusack, J. J. (2018). Camera trapping in Africa: Paving the way for ease of use and consistency. *African Journal of Ecology*, *56*(4), 690–693. https://doi.org/10.1111/aje.12581

Bahaa-el-din, L., Henschel, P., Aba'a, R., Abernethy, K., Bohm, T., Bout, N., Coad, L. M., Head, J., Inoue, E., Lahm, S. A., & Lee, M. (2013). Notes on the distribution and status of small carnivores in Gabon. *Small Carnivore Conservation*, *48*, 19–29.

Beery, S., Morris, D., Yang, S., Simon, M., Norouzzadeh, A., & Joshi, N. (2019). Efficient pipeline for automating species ID in new camera trap projects. *Biodiversity Information Science and Standards*, *3*, e37222. https://doi.org/10.3897/biss.3.37222

Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 456–473). Retrieved from https://openaccess.thecvf.com/content_ECCV_2018/html/Beery_Recognition_in_Terra_ECCV_2018_paper.html

Bessone, M., Kühl, H. S., Hohmann, G., Herbinger, I., N'Goran, K. P., Asanzi, P., Costa, P. B. D., Dérozier, V., Fotsing, E. D. B., Beka, B. I., Iyomi, M. D., Iyatshi, I. B., Kafando, P., Kambere, M. A., Moundzoho, D. B., Wanzalire, M. L. K., & Fruth, B. (2020). Drawn out of the shadows: Surveying secretive forest species with camera trap distance sampling. *Journal of Applied Ecology*, *57*(5), 963–974. https://doi.org/10.1111/1365-2664.13602

Borchers, D. L., & Efford, M. G. (2008). Spatially explicit maximum likelihood methods for capture-recapture studies. *Biometrics*, *64*(2), 377–385. https://doi.org/10.1111/j.1541-0420.2007.00927.x

Cardoso, A. W., Malhi, Y., Oliveras, I., Lehmann, D., Ndong, J. E., Dimoto, E., Bush, E., Jeffery, K., Labriere, N., Lewis, S. L., White, L. T. J., Bond, W., & Abernethy, K. (2020). The role of forest elephants in shaping tropical forest–savanna coexistence. *Ecosystems*, *23*(3), 602–616. https://doi.org/10.1007/s10021-019-00424-3

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., & White, E. P. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), 1424–1432. https://doi.org/10.1073/pnas.1710231115

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

Fiske, I., & Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, *43*(10), 1–23.

Glover-Kapfer, P., Soto-Navarro, C. A., & Wearn, O. R. (2019). Camera-trapping version 3.0: Current constraints and future priorities for development. *Remote Sensing in Ecology and Conservation*, *5*(3), 209–223. https://doi.org/10.1002/rse2.106

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *ArXiv:1706.04599 [Cs]*. Retrieved from http://arxiv.org/abs/1706.04599

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision – ECCV 2016* (pp. 630–645). Springer International Publishing. https://doi.org/10.1007/978-3-319-46493-0_38

Howard, J., & Gugger, S. (2020). Fastai: A layered API for deep learning. *Information*, *11*(2), 108. https://doi.org/10.3390/info11020108

Kuhn, M. (2020). *caret: Classification and regression training*. R Package Version 6.0-86. Retrieved from https://CRAN.R-project.org/package=caret

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. *ArXiv:1607.02533 [cs, Stat]*. Retrieved from http://arxiv.org/abs/1607.02533

Lucas, T. C. D., Moorcroft, E. A., Freeman, R., Rowcliffe, J. M., & Jones, K. E. (2015). A generalised random encounter model for estimating animal density with remote sensor data. *Methods in Ecology and Evolution*, *6*(5), 500–509. https://doi.org/10.1111/2041-210X.12346

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, *84*(8), 2200–2207. https://doi.org/10.1890/02-3090

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248–2255. https://doi.org/10.1890/0012-9658(2002)083%5B2248:ESORWD%5D2.0.CO;2

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., & Clune, J. (2019). A deep active learning system for species identification and

counting in camera trap images. *ArXiv:1910.09716 [Cs, Eess, Stat]*. Retrieved from http://arxiv.org/abs/1910.09716

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 of the United States of America*, *115*(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

O'Brien, T. G., Ahumada, J., Akampurila, E., Beaudrot, L., Boekee, K., Brncic, T., Hickey, J., Jansen, P. A., Kayijamahe, C., Moore, J., Mugerwa, B., Mulindahabi, F., Ndoundou-Hockemba, M., Niyigaba, P., Nyiratuza, M., Opepa, C. K., Rovero, F., Uzabaho, E., & Strindberg, S. (2020). Camera trapping reveals trends in forest duiker populations in African National Parks. *Remote Sensing in Ecology and Conservation*, *6*(2), 168–180. https://doi.org/10.1002/rse2.132

Orbell, C., & Whytock, R. C. (2021). Datasets for Robust ecological analysis of camera trap data labelled by a machine learning model. DataSTORRE: Stirling Online Repository for Research Data. Retrieved from http://hdl.handle.net/11667/170

Rowcliffe, J. M. (2019). *activity: Animal activity statistics. R Package v 1.3*. Retrieved from https://CRAN.R-project.org/package=activity

Rowcliffe, J. M., Kays, R., Kranstauber, B., Carbone, C., & Jansen, P. A. (2014). Quantifying levels of animal activity using camera trap data. *Methods in Ecology and Evolution*, *5*(11), 1170–1179. https://doi.org/10.1111/2041-210X.12278

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*(4), 835–841. https://doi.org/10.1890/0012-9658(2006)87%5B835:GSOMAF%5D2.0.CO;2

Schneider, S., Greenberg, S., Taylor, G. W., & Kremer, S. C. (2020). Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and Evolution*, *10*(7), 3503–3517. https://doi.org/10.1002/ece3.6147

Schneider, S., Taylor, G. W., & Kremer, S. C. (2018). Deep learning object detection methods for ecological camera trap data. *ArXiv:1803.10842 [Cs]*. Retrieved from http://arxiv.org/abs/1803.10842

Schneider, S., Taylor, G. W., Linquist, S., & Kremer, S. C. (2019). Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, *10*(4), 461–470. https://doi.org/10.1111/2041-210X.13133

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 – Learning rate, batch size, momentum, and weight decay. *ArXiv:1803.09820 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1803.09820

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015). Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, *2*(1), 150026. https://doi.org/10.1038/sdata.2015.26

Świeżewski, J., & Whytock, R. C. (2021). Public release for archiving to accompany Whytock & Swiezewski et al 2021. Robust ecological analysis of camera trap data labelled by a machine learning model. *Zenodo*. http://doi.org/10.5281/zenodo.4534108

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., … Miller, R. S. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, *10*(4), 585–590. https://doi.org/10.1111/2041-210X.13120

Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking model scaling for convolutional neural networks. *ArXiv:1905.11946 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1905.11946

Wei, W., Luo, G., Ran, J., & Li, J. (2020). Zilong: A tool to identify empty images in camera-trap data. *Ecological Informatics*, *55*, 101021. https://doi.org/10.1016/j.ecoinf.2019.101021

Whytock, R. C. (2021). R code to accompany Whytock and Swiezewski et al 2021. Robust ecological analysis of camera trap data labelled by a machine learning model. *Zenodo*, https://doi.org/10.5281/zenodo.4486332

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*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

Xie, S., Girshick, R., Dollar, P., Tu, Z., & He, K. (2017). *Aggregated residual transformations for deep neural networks* (pp. 1492–1500). Retrieved from https://openaccess.thecvf.com/content_cvpr_2017/html/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.html

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., & Smola, A. (2020). *ResNeSt: Split-attention networks*. Retrieved from http://arxiv.org/abs/2004.08955

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Whytock RC, Świeżewski J, Zwerts JA, et al. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods Ecol Evol*. 2021;12:1080–1092. https://doi.org/10.1111/2041-210X.13576