Received: 25 May 2018 | Accepted: 20 September 2018

DOI: 10.1111/2041-210X.13099

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Identifying animal species in camera trap images using deep learning and citizen science

Marco Willi[1] | Ross T. Pitman[2,3] | Anabelle W. Cardoso[4] | Christina Locke[5] | Alexandra Swanson[6] | Amy Boyer[7] | Marten Veldthuis[6] | Lucy Fortson[1]

[1]School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota; [2]Panthera, New York, New York; [3]Department of Biological Sciences, Institute for Communities and Wildlife in Africa, University of Cape Town, Cape Town, South Africa; [4]School of Geography and the Environment, University of Oxford, Oxford, UK; [5]Wisconsin Department of Natural Resources, Office of Applied Science, Madison, Wisconsin; [6]Department of Astrophysics, University of Oxford, Oxford, UK and [7]Adler Planetarium, Chicago, Illinois

Correspondence
Marco Willi
Email: will5448@umn.edu

Funding information
National Science Foundation, Grant/Award Number: IIS 1619177

Handling Editor: Oscar Gaggiotti

Abstract

1. Ecologists often study wildlife populations by deploying camera traps. Large datasets are generated using this approach which can be difficult for research teams to manually evaluate. Researchers increasingly enlist volunteers from the general public as citizen scientists to help classify images. The growing number of camera trap studies, however, makes it ever more challenging to find enough volunteers to process all projects in a timely manner. Advances in machine learning, especially deep learning, allow for accurate automatic image classification. By training models using existing datasets of images classified by citizen scientists and subsequent application of such models on new studies, human effort may be reduced substantially. The goals of this study were to (a) assess the accuracy of deep learning in classifying camera trap data, (b) investigate how to process datasets with only a few classified images that are generally difficult to model, and (c) apply a trained model on a live online citizen science project.
2. Convolutional neural networks (CNNs) were used to differentiate among images of different animal species, images of humans or vehicles, and empty images (no animals, vehicles, or humans). We used four different camera trap datasets featuring a wide variety of species, different habitats, and a varying number of images. All datasets were labelled by citizen scientists on Zooniverse.
3. Accuracies for identifying empty images across projects ranged between 91.2% and 98.0%, whereas accuracies for identifying specific species were between 88.7% and 92.7%. Transferring information from CNNs trained on large datasets ("transfer-learning") was increasingly beneficial as the size of the training dataset decreased and raised accuracy by up to 10.3%. Removing low-confidence predictions increased model accuracies to the level of citizen scientists. By combining a trained model with classifications from citizen scientists, human effort was reduced by 43% while maintaining overall accuracy for a live experiment running on Zooniverse.
4. Ecology researchers can significantly reduce image classification time and manual effort by combining citizen scientists and CNNs, enabling faster processing of data from large camera trap studies.

KEYWORDS
animal identification, camera trap, citizen science, convolutional neural networks, deep learning, machine learning

1 | INTRODUCTION

Ecologists often deploy camera traps to study wildlife populations in areas of interest (O'Connell, Nichols, & Karanth, 2011). Such cameras are remote, independent devices, triggered by motion and infrared sensors that provide researchers with images of passing animals (e.g., see Figure 1). After collection, images have to be classified according to the study's goals to produce useful ecological data for analysis. Recent studies encompass millions of images (e.g., Swanson et al., 2015), which is too many for research teams to manually process and classify within a reasonable time-frame. Citizen science is a valuable approach in such cases, as it enables harnessing the collective effort of volunteers from the general population (citizen scientists). Citizen scientists have become essential contributors to general science, ecology, and camera trap projects (Dickinson, Zuckerberg, & Bonter, 2010; Silvertown, 2009; Swanson et al., 2015). Species identification leveraging the collective labour of thousands of citizen scientists can result in accurate classifications (Swanson, Kosmala, Lintott, & Packer, 2016) while reducing the time to process large datasets.

1.1 | Citizen science and camera trap projects

Zooniverse (www.zooniverse.org) is the largest online citizen science platform and provides researchers with access to millions of volunteers (Fortson et al., 2012; Swanson et al., 2015; Zevin et al., 2016). The platform allows researchers to build a custom project website (using the "project builder", www.zooniverse.org/lab) and to supply guidelines, additional information, classification data, and annotation tools for citizen scientists (see Supporting Information for an example). Research teams define workflows, which comprise specific tasks the citizen scientists are asked to perform, for example, identifying animals in camera trap images. Zooniverse has hosted 147 public projects as of May 2018 of which 29 were camera trap projects. Collectively, the camera trap projects contributed over 17 million identification tasks for which citizen scientists submitted more than 63 million classifications. The first camera trap project (Snapshot Serengeti) was launched in 2012, whereas 12 new camera trap projects have been published between January and May 2018 alone. During this time the total number of projects on Zooniverse has exploded by a factor of 7.4 to 147, whereas the number of registered volunteers has only grown by a factor of 2.4 to over 1.7 million. This has increased the competition for volunteers and is leading to increasing delays between data collection and subsequent analyses.

Further concerns for some camera trap projects include the need to remove sensitive images before publishing them on citizen science platforms, for example, images of humans for privacy reasons or images of rare species to prevent exposing their location to poachers. Additionally, some camera trap projects contain a large number of empty images (containing no animals), caused when camera traps are triggered by insignificant objects like overgrown vegetation moving in the wind. Research teams often aim at identifying and removing empty images as efficiently as possible, for example, Hines, Swanson, Kosmala, and Lintott (2015).

To address these issues, automatic classification of camera trap images has been a focus of research in computer vision and machine learning. Recent advances in using techniques from deep learning have enabled researchers to improve automatic species identification significantly (Gomez Villa, Salazar, & Vargas, 2017; Norouzzadeh et al., 2018).

1.2 | Machine learning and image classification

In machine learning, particularly supervised learning, the goal is to learn how to map inputs (data) to outputs of interest by applying specific learning algorithms. In the case of image classification, the aim is to train an algorithm which can process and label images with classes of interest (e.g., animal species). In recent years, image classification has been dominated by convolutional neural networks (CNNs) as demonstrated in recent ImageNet Large Scale Visual Recognition Challenges (ILSVRC) (Krizhevsky, Sutskever, & Hinton, 2012; Russakovsky et al., 2015).

CNNs (LeCun et al., 1989) consist of two linked main components, a convolutional part which extracts local features from images and a fully connected part which maps the learned features to outputs. Unlike earlier methods, CNNs do not rely on manually defined features. Instead, the network learns spatial features by updating its parameters (weights) during model training by the propagation of errors from the output towards the input. The exact structure of a CNN (the sequence of operations applied to the data by the network) is known as its architecture. Figure 2 schematically shows the architecture of a CNN and its main module (a layer) which consists of filters which are convolved over the input (to measure their spatial occurrence), activation functions, and pooling (sub-sampling) operations. A layer normally results in smaller feature maps which are then passed to the next layer. Typically, many such layers are stacked sequentially allowing for complex feature extraction. The number of layers in an architecture is referred to as the depth of a network, thus, deep learning implies neural networks with many layers.

CNNs often are very complex and can contain millions of parameters which all have to be learned from training data. The more complex a model, that is, the more parameters it has, the more training data are usually required to learn suitable parameter values. A lack

FIGURE 1 Examples of camera trap images from the different projects used in this study. From left to right: Snapshot Serengeti, Camera CATalogue, Elephant Expedition, and Snapshot Wisconsin

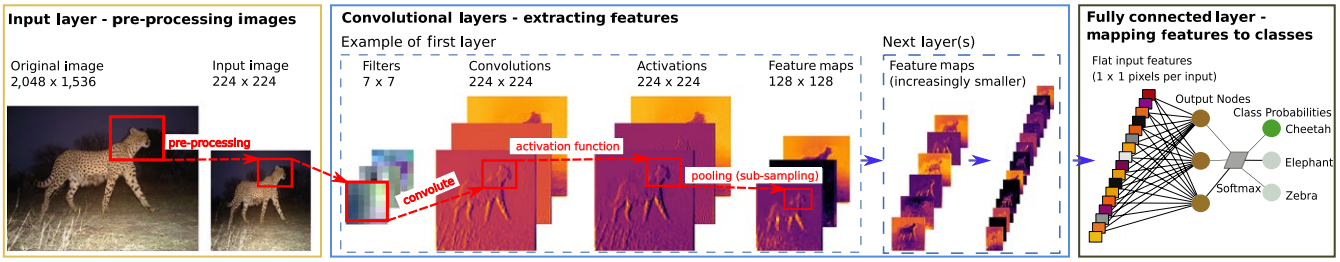

FIGURE 2 Schematic illustration of a CNN architecture. Each of (typically) many convolutional layers in a CNN learns (small) filters which are convoluted over the input from the previous layer to detect their spatial occurrence. Subsequently, after applying an activation function (allowing for nonlinear relationships), new feature maps are obtained by down sampling techniques (pooling) which are then passed as input to the next layer. Finally, a fully connected layer connects the final ("flat") feature maps to the output classes via a softmax transformation which converts the output node values to class probabilities. The red square tracks a region of the depicted input image through the first convolutional layer showing real visualizations by applying a filter which detects edge-like structures

of training data may lead to overfitting, meaning that a model does not generalize well to new data (i.e., the model may focus on specific patterns in the training data not relevant for newly gathered data). Besides having enough training data, additional strategies to avoid overfitting are usually necessary during model training (Glorot & Bengio, 2010; Krizhevsky et al., 2012). One strategy for preventing overfitting, especially suited for small datasets, is transfer-learning (see section 2.4), which we applied on different sizes of training datasets to evaluate its performance.

1.3 | Automatic animal identification

Initial research in automatic animal identification focused on matching species-specific patterns in images and required a substantial amount of manual preprocessing. Even so, the accuracies that were achieved, for example, 82% by Yu et al. (2013), were not comparable to human accuracies of 96.6% (Swanson et al., 2016). More recent studies in automatically identifying animal species using CNNs have reported accuracies around 90% with some manual preprocessing (Gomez Villa et al., 2017) or more complex pipelines involving automatic preprocessing (Giraldo-Zuluaga, Salazar, Gomez, & Diaz-Pulido, 2017). The most recent advances by Norouzzadeh et al. (2018) have reported accuracies of 93.8% and have matched human accuracy on over 99% of all images.

Our study aims at applying and validating the use of CNNs on a larger variety of camera trap datasets as compared to previous studies. Norouzzadeh et al. (2018) reported impressive accuracies using the Snapshot Serengeti dataset which comprises 3.2 million images (Swanson et al., 2015)—much more than most camera trap datasets as observed on Zooniverse. Training effective image classification models often requires large datasets (e.g., the famous ImageNet dataset contains 1.2 million images) and thus makes that approach feasible for datasets of the size of Snapshot Serengeti. To investigate the applicability of CNNs in more realistic (smaller) scenarios, we included several smaller datasets each with significantly fewer than one million images.

Furthermore, using transfer-learning, our work investigates how to leverage models trained on large camera trap datasets to smaller datasets. This technique is well-known and was applied in previous studies (Gomez Villa et al., 2017; Norouzzadeh et al., 2018), however, by transferring models trained on noncamera trap datasets (e.g., ImageNet) which is expected to be less effective than transferring from models trained to accomplish a similar task (i.e., animal identification). We transferred data from a model trained on a more recent version of the Snapshot Serengeti dataset comprising 7.3 million images. Transfer-learning could help citizen science platforms like Zooniverse to more quickly train and use models for new datasets with few labelled images.

Finally, our study is the first to demonstrate and evaluate a use case of how trained models can be applied to classify new camera trap images by combining human annotations and model classifications in real time on an online citizen science platform. Previous studies, for example, Norouzzadeh et al. (2018), estimated the reduction in human effort for classifying new images by extrapolating model performance as measured on historical data and without combining human annotations with model predictions.

TABLE 1 Overview and comparison of all datasets used in this study

	Snapshot Serengeti	Camera CATalogue	Elephant Expedition	Snapshot Wisconsin
Images total (millions)—used in empty/species models	7.3	0.52	0.42	0.5
	1.1/1.2	0.47/0.14	0.12/0.05	0.28/0.3
Empty/vehicle (%)	81.8/-	21.7/47.8	83.9/-	31.5/-
Species (total/modelled)	54/51	55/48	9/9	45/31
Location	Tanzania	South Africa	Gabon	USA
Images per capture event	3	1	1	3
Trigger block time (s)	60	8–15	1	15
Vegetation	Savanna	Savanna	Forest–savanna mosaic	Forest (66%)
Number of cameras	225	750	40	1037

1.4 | Goals of this study

1. To assess the performance of CNNs for classifying images into different animal species, empty images, and human or vehicle images from four camera trap projects.
2. To assess how an established technique called "transfer-learning" can help train models for projects with a low number of (human classified) images which are generally more challenging to accurately model.
3. To implement and evaluate a trained model in the context of a live citizen science project on Zooniverse to observe its applicability under real-world conditions and test for improvements in classification efficiency.

2 | DATASETS AND METHODS

2.1 | Datasets

We used four camera trap datasets in this study (see Table 1 for overview and Supporting Information for descriptions) which were collected by different research teams. Each dataset consisted of camera trap images and their annotations provided by citizen scientists on Zooniverse. The images of three datasets were collected in Africa: Snapshot Serengeti (SS), Camera CATalogue (CC), and Elephant Expedition (EE), whereas images from Snapshot Wisconsin (SW) were collected in North America. These datasets differed in aspects such as dataset size, camera placement, camera configuration, and species coverage which allows for drawing more general conclusions. Furthermore, the SS dataset was used in previous machine learning studies, thus facilitating detailed comparisons of model results.

The datasets featured between nine and 55 species and exhibited significant imbalance in how often different species were photographed: Pearson's moment coefficient of skewness ranged between 1.58 and 5 (see Supporting Information for species distributions). Extremely rare species (photographed <27 times) were excluded from the study in training the models to ensure images of a specific species were present in all the dataset splits (training, validation and test) and thus is an arbitrary limit chosen for this study (Technically, a model can be trained to identify any species when there is at least one image available featuring that species [by only including it in the training set]). Species overlap between the different projects was often high: Of the 51 species we modelled in the SS dataset 35 were also present in the CC dataset. Five of the nine species from the EE dataset were also covered by the SS and CC datasets. Unsurprisingly, SW had no species overlap with the other projects. Variabilities in vegetation density and camera placement impacted camera field of vision and the frequency of photographs triggered by moving vegetation or human activity. The share of empty images ranged between 21.7% and 83.9%, whereas human triggered images contributed between close to 0% and 47.8%.

When a camera trap is triggered we refer to it as a capture event. Camera configurations defined how many photographs were taken and how long the camera trigger was blocked after a capture event. Regardless of how many photographs a capture event comprised, volunteers always annotated capture events and not individual images. During model training, however, the model was provided with individual images and was unaware of which images belonged to which capture event (all images of a dataset were repeatedly and randomly shown to the model during training). The images were labelled according to the label inherited from their capture event. The images in all datasets were of high quality with resolutions ranging between 800 × 600 and 2,048 × 1,436 pixels.

2.2 | Citizen science and image classification

Each capture event was annotated by multiple citizen scientists using the Zooniverse platform. The research teams for each project configured how many annotations per capture event they required before it was considered finished (retired)—balancing classification quality and time to process the whole dataset. The retirement limits were class specific and often dynamically adjusted based on volunteer consensus. For example, SW collected annotations for a given capture event until seven volunteers agreed on one species with a cap at 15 annotations. Overall, the retirement limits ranged between

one (images of humans for SW and CC) and 25 annotations (no consensus in the retirement procedure of SS).

The researchers built project-specific workflows by selecting different options for tasks to be done by the citizen scientists. For the SS, EE, and SW datasets the researchers used the "survey task" workflow in which the capture events were shown to citizen scientists along with all possible classifications to choose from. CC used a novel "cascade filtering" procedure to improve the retirement speed of empty and vehicle images. In this latter approach, images were initially processed through the "Empty or Not" workflow asking volunteers to identify empty images. Nonempty images were subsequently routed to the "Vehicle or Not" workflow aimed at removing vehicle images. The remaining images were then passed to a "survey task" workflow for species identification. This "cascading" of easier, binary workflows required less overall classifications and may lead to faster decision making while annotating images that collectively reduce the overall processing time.

Each capture event was annotated by multiple volunteers potentially leading to conflicting and/or multiple labels. Because our model required each image to be associated with one unique class, we aggregated multiple annotations. We used the plurality algorithm as described in Swanson et al. (2015) to achieve consistent labelling across all projects. The algorithm first determines the number of distinct species that each image contains. If the majority of volunteers identified more than one species in a capture event we excluded it from the study due to the additional complexity involved in modelling such images (multi-label classification) and the low occurrence of such situations (<1% for all projects). For the remaining images the most frequently annotated species was selected as the final label.

2.3 | Model training

2.3.1 | Model architecture and parameters

To balance accuracy and computational costs, all models were trained using the ResNet18 model architecture. ResNet18 is a less complex version of ResNet152, the latter of which won the ILSVRC in 2015 (He, Zhang, Ren, & Sun, 2016). Comparisons of a variety of state-of-the art CNN architectures on an earlier version of the SS dataset by Norouzzadeh et al. (2018) showed that ResNet18 is only 0.5% less accurate than the best model evaluated by them. The model contains over 11 million parameters, has 18 convolutional layers and is smaller than many other state-of-the art models, thus reducing training time. To train the models we used Tensorflow (Abadi et al., 2016), an open-source graph computation framework, and servers hosted on Amazon Web Services (AWS) using one GPU with 12 GB of RAM. Details about parameter choices and image preprocessing can be found in the Supporting Information.

2.3.2 | Data splitting

We trained independent models for each dataset by randomly assigning all capture events of a dataset, while preserving class distributions, to one of three sets: a training set (90% of the data), a validation set (5% of the data), and a test set (5% of the data). The models were trained on the training set, while being monitored on the validation set to reduce overfitting. This was achieved by finishing model training when accuracy stopped improving on the validation set. Final results were reported using the test set. For more details regarding data splitting see the Supporting Information.

Since we split the datasets randomly, images from a specific camera trap could be in any of the three splits, therefore, the model might have learned specific features or species biases for individual locations. Thus, the reported accuracies resemble the performance that could be obtained if we were to continue collecting and classifying data using the same camera trap network. It is possible that model performance would be lower for images collected from camera traps the model has not encountered during model training.

2.3.3 | Two models: identifying empty images and species

For each dataset we trained two different models. One model differentiated between empty and nonempty images, whereas the second model differentiated between species. For the CC dataset the first model also included vehicles as a separate class. The reasons for training two different models were: (a) to reduce the likely biases models would have towards the majority class, which was expected to be significant as the share of empty images was generally high, and (b) to more closely resemble the application of the model in the live experiment ("cascade filtering") aimed at increasing accuracy by more closely learning the target task.

The reported results for the postprocessed datasets do not consider applying the two models sequentially. This introduced a positive bias to the species models as they were evaluated on images we were certain contained an animal, that is, the model could not make the mistake of predicting a species when the image was empty. Instead, the evaluation of the CC live experiment (see section 2.6) did consider this sequence and thus resembled a realistic scenario.

We did not directly (i.e., without transfer-learning) apply models trained on one dataset to another dataset even if the identification of empty images is the same task. Due to the substantial differences in vegetation, study sites, camera placement, and species we did not expect our models to generalize well to other datasets. We believe, however, this is an interesting and fruitful area of research.

2.4 | Transfer-learning

In order to investigate the possibilities of achieving high accuracies on smaller training datasets, we implemented the transfer-learning technique. Transfer-learning (Yosinski, Clune, Bengio, & Lipson, 2014) involves copying the weights learned on a base model to a target model. This potentially improves model accuracies, reduces model training time, and reduces the amount of labelled data required. To distinguish how we trained our models we refer to training without transfer-learning as training from scratch.

For transfer-learning the base models used in this study were the models trained on the SS dataset. Those models were trained on the largest dataset and featured a wide variety of species and thus were expected to learn features represented in the other projects. We applied transfer-learning by copying and freezing all the weights from the convolutional layers (feature extraction) and only re-learned the weights of the fully connected layers (see Figure 2). Initial experiments did not show an additional benefit of re-learning weights of convolutional layers (The accuracy of the CC species model was 1.5% lower when allowing for adaptation of all weights, the accuracy of the EE species model was 0.2% lower. We point out that different model training parameters could influence these results).

2.5 | Model output and confidence thresholding

The CNN outputs probabilities over all classes for each image, that is, if we train a model to distinguish between 20 classes it will output 20 probabilities for each image. To assign a class to a particular capture event, we assigned the class with the highest probability across all images of the capture event. This probability can be interpreted as a confidence measure in the model's prediction; the higher the predicted probability, the more confident the prediction. This allowed us to apply confidence thresholding which ignores the model's opinion if its highest probability was below a certain threshold. By ignoring low-confidence predictions the overall accuracy of the model can be increased while reducing coverage (the share of images for which we considered the model prediction).

2.6 | Model application on a live zooniverse project

While the above methods describe how we modelled camera trap data that has already been recorded and annotated, we also studied the impact of applying a trained model on newly collected images in a live project on Zooniverse.

Our goal was to decrease the number of volunteer annotations required by considering the output of a trained model for a camera trap project on Zooniverse. We conducted an experiment where machine predictions were combined with human annotations for a new batch of CC images, which was not used in the rest of this study. To evaluate the experiment we assessed the efficiency gain and the impact on the quality of the final aggregated labels.

We randomly assigned 50% of all capture events to an experiment group, retaining the rest as a control group, and implemented new image retirement thresholds for this experiment. In the (nonexperiment) "cascade filtering" workflow of CC, capture events in the "Empty or Not" and "Vehicle or Not" workflows were retired if the first two volunteers agreed with each other (more otherwise), whereas five volunteer annotations were required for the species workflow, regardless of agreement. For the experiment, we altered the logic such that if the first volunteer agreed with the model, the capture event was retired if it was in the "Empty or Not" or "Vehicle or Not" workflow. For the "species" workflow, an image was retired if the first two volunteers agreed with the model. In cases where there was no agreement with the model, the capture event fell back to the nonexperiment logic. Images were excluded from the experiment where the model's top prediction was a rare species (fewer than 2,000 images in the dataset), a species with low historical accuracy (<90%), or resulted in a model prediction below 85% (only for species predictions). After these exclusions 17 species remained in the dataset (totalling over 88% of all species images as classified by the model) plus all predicted empty and vehicle images. To classify the images, the empty model was applied first and the species model was subsequently applied to the images classified as containing a species by the first model (see discussion in section 2.3.3).

To evaluate the classification quality we simulated the application of the experiment logic on the control group and compared the labels of the simulation with the effective labels as annotated by the citizen scientists. To evaluate the efficiency gain we measured the reduction in annotations until all capture events were classified. Therefore, the sole utility of the experiment group was to show the new logic could be implemented technically while still profiting from efficiency gains on a live project. The control group was used for the quantitative evaluation. (Due to the random splitting into control-and experiment group we can reasonably assume that the results of the control group are representative of the experiment group)

3 | RESULTS

To evaluate our models, we report the accuracy as the percentage of correct model predictions for camera trap events compared to the aggregated volunteer opinion (ground truth). To put the reported accuracies in context: Citizen scientists classifying species images from Serengeti National Park reached an aggregated accuracy of 96.6% (Swanson et al., 2016) as compared to expert opinion. See Supporting Information for results and a discussion regarding statistical uncertainties. Generally, the width of the 95% confidence intervals ranged between 0.3% and 2.0% and were inversely correlated with the size of the modelled datasets.

3.1 | Model accuracies—training from scratch vs. transfer-learning

Model accuracies for all trained models are shown in Figure 3. Transfer-learning outperformed training from scratch on all datasets for the species models. Even the SW dataset substantially profited from transfer-learning, although there was no species overlap and the vegetation is quite different as compared to the SS dataset. Similarly, the models identifying empty images profited from transfer-learning with the exception of the CC dataset (0.3% lower accuracy - statistically significant). Identifying empty images consistently outperformed the species models by 2.4%–5.9% when trained from scratch and by 0%–5.4% when trained with transfer-learning.

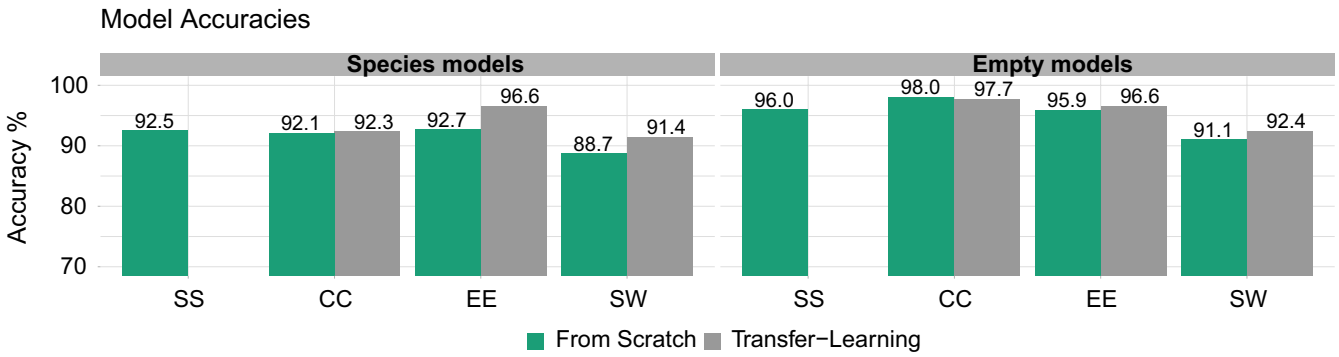

FIGURE 3 Shown are the average model accuracies (%) for all datasets of the species models (left panel) and the empty models (right panel)—by training the models from scratch (green) and by transfer-learning (grey). Transfer-learning results for SS are missing because it was used as the base model for transfer-learning

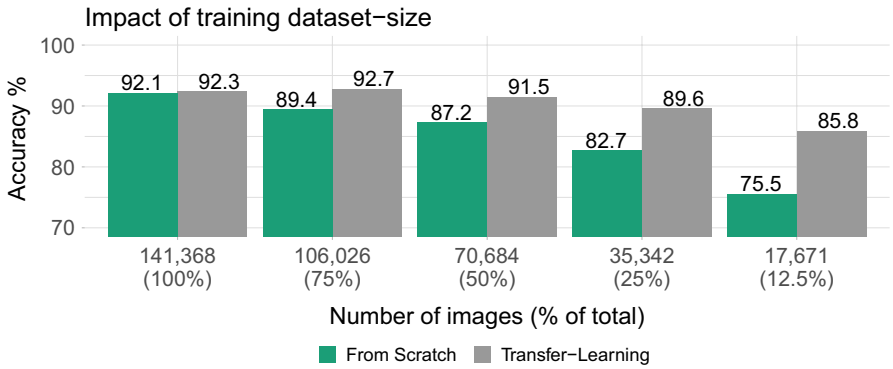

FIGURE 4 Accuracies of the CC species model trained on different training dataset sizes with either training from scratch (green) and transfer-learning (grey). Note that the model using 100% of the data performed slightly worse than the model using 75% of the data when using transfer-learning. This is unexpected but possible considering the random factors involved (see discussion in Supporting Information). The difference is not statistically significant

We assessed how transfer-learning and training from scratch were impacted by dataset size by selecting increasingly smaller training splits of the CC dataset (see Figure 4). The results show the increasing benefit of transfer-learning for smaller training (annotated) datasets. While transfer-learning and training from scratch yielded similar accuracy when using the full training dataset with 141,368 images, the difference increased to over 10% when using only 17,671 images (12.5% of the training data). Using only 70,684 images (50% of the training data) yielded an accuracy reduction of <1% using transfer-learning, and led to a 4.9% reduction when trained from scratch. Note that the observed reductions in accuracies are specific to this dataset (see also results at end of section 3.3 and discussion in section 4).

Selected examples of misclassified images are shown in Figure 5. In these cases, the confidence indicated by the model is displayed by the top bars below each image and is clearly below 95%. Visual inspection of misclassified predictions revealed difficulties when there was a low contrast between animal and background, for example, in night-time images, when an animal was very close to the camera, or if the camera's flash or sun flares obscured the animal. Also, the models generally had difficulties with rare species, even if they were well imaged.

3.2 | Confidence thresholding

With confidence thresholding, predictions were ignored if the model assigned low probabilities and thus indicated uncertainty. Figure 6 shows the effect of confidence thresholding on prediction accuracy and dataset coverage for different thresholds on the EE species model trained from scratch (see all figures in the Supporting Information). Accuracy for all models increased with increasing thresholds while reducing coverage. However, even if we set the threshold to 99%, which resulted in accuracies between 96.7% and 98.9%, coverage remained at 76%–86.5% for the species models. The empty models reached slightly higher accuracies (97.5%–99.4%) and coverage (73.7%–94.7%). Noteworthy are the high accuracies of the empty models for CC (99.4%) and EE (98.9%), as well as accuracies over 99% for several frequent species across all datasets (ignoring statistically insignificant accuracies for rare species).

3.3 | Accuracies and recall for individual species

Accuracies of individual species varied strongly and were positively correlated with the amount of training data available for a specific species. For example, on the SS dataset, when the model classified

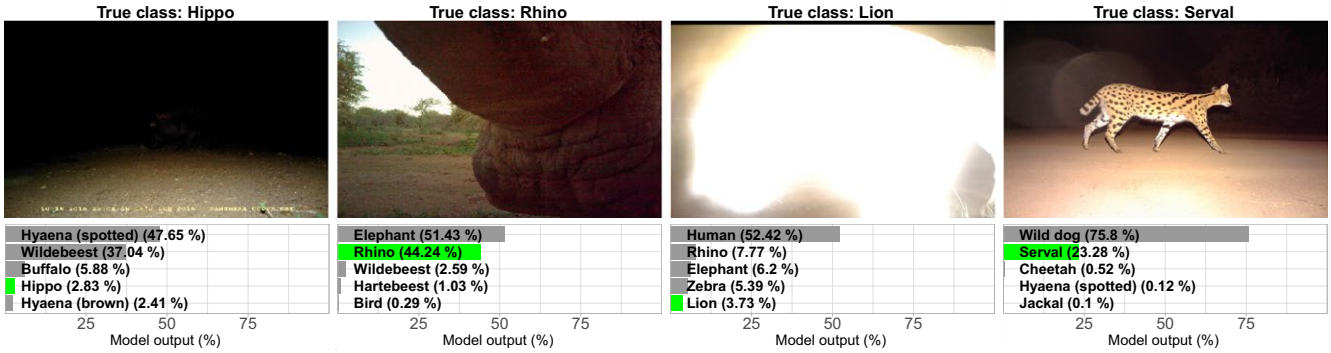

FIGURE 5 Examples of misclassified images from the CC dataset, highlighting some of the challenges. From left to right: A night-time image of an animal with very low contrast to the background, a rhino standing very close to the camera, an overexposed lion, and a rare serval. Beneath each image, the top five predictions of the model are shown. All true classes were derived from annotations by citizen scientists

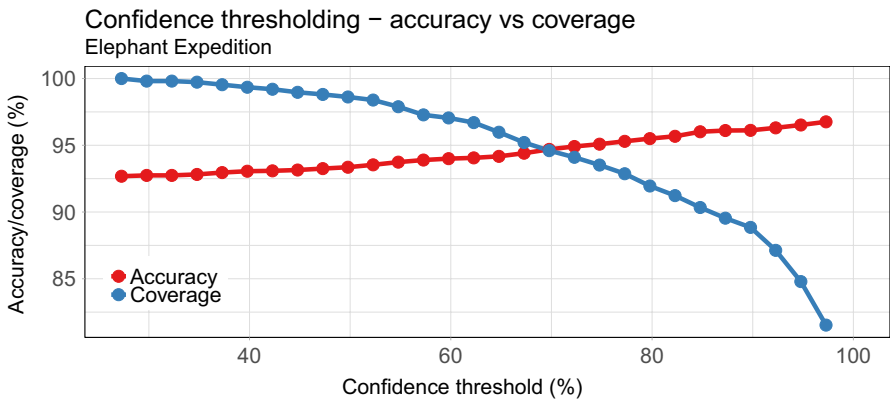

FIGURE 6 Effect of applying different confidence thresholds (*x* axis) on accuracy (*y* axis, red line) and coverage (*y* axis, blue line) for the EE species model. Note zero suppression of *y*-axis

a capture event as an extremely rare striped hyena it was always wrong (only one such case), whereas accuracies for more common species like zebra and wildebeest were around 97% (see Figure 7 for selected species accuracies of SS and Supporting Information for accuracies for all datasets and species). Confidence thresholding had a larger effect on rarer species as shown in Figure 7. For example, high confidence predictions of the rare ostrich class were always correct (27 such cases).

A binomial regression showed that the number of images in the training set, the relative occurrence in the training set, and the particular species were statistically significant ($p < 0.05$) factors in explaining the variation in observed species accuracies. Median accuracy for species with 1,000–10,000 images in the training set was 77% and 91% with confidence thresholding (threshold at 95%). Species occurring more than 10,000 times in the training set had a median accuracy of 90% and 97% with confidence thresholding (threshold at 95%).

Figure 8 shows the proportion of capture events of a specific class (ground truth) which were correctly classified by the model (referred to as recall) for selected species of the SS dataset (see Supporting Information for all recall data). Recall was strongly correlated with species frequency and ranged between 93.3% and 97.6% across all datasets for the most frequent species and was 0% for some of the rare species. Generally, confidence thresholding reduces recall. The observed reductions were stronger for rare species.

3.4 | Live experiment

3.4.1 | Efficiency gain

To process all experiment-eligible capture events, 49% fewer annotations by citizen scientists were required. Accounting for the noneligible subjects for which the model supplied no opinion and thus were processed by the standard logic, 43% fewer annotations were needed to retire all images.

3.4.2 | Classification quality

Only 0.22% of all capture events had a different final label due to the newly defined logic (Accordingly, the accuracy of the human-model combination was 99.78%. Using only model predictions would have resulted in an accuracy of 93.4%). Visual inspection showed that most of these cases were either due to difficulty spotting obscured animals by the first volunteer (and the model), or concurrence of humans and vehicles in an image, causing confusion by volunteers whether to choose the human or vehicle class. In total 91.2% of all images were retired with fewer annotations because the first volunteers agreed with the model. In 8.6% of the cases, the machine and the first volunteers disagreed, and the standard logic was applied for retirement. However, in 47% of the disagreements, the final plurality label matched the model label, indicating the first volunteer(s) were incorrect about half the time when there was a disagreement.

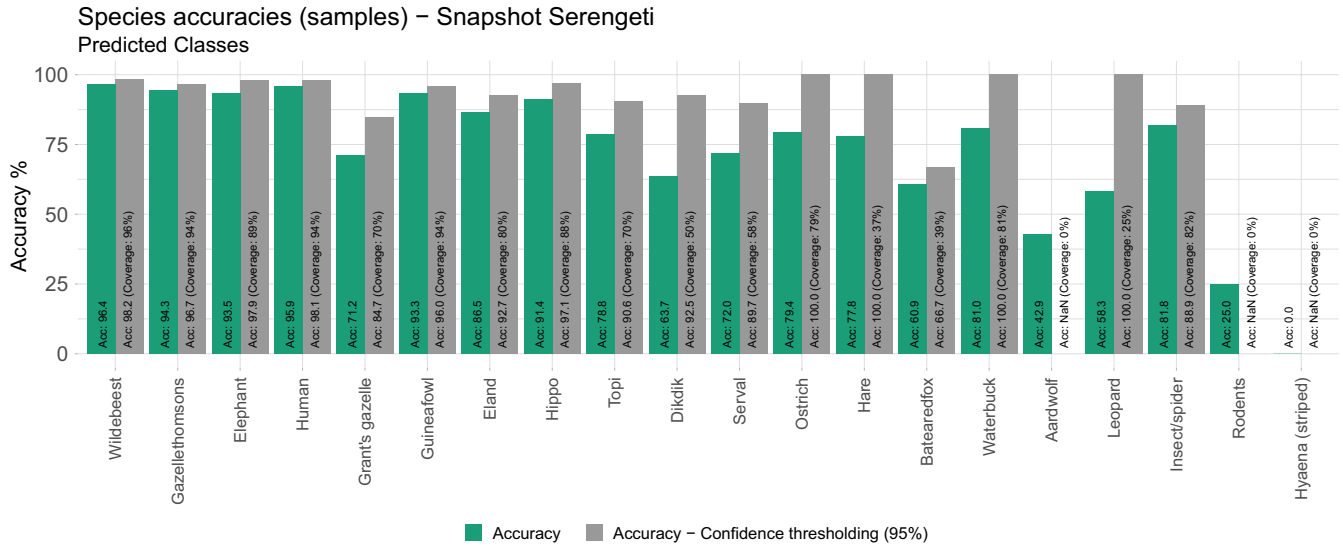

FIGURE 7 Species accuracies for the SS species model (only a selection of species are shown)—with all capture events (green) and only high confidence events (grey, 95% confidence or higher). Exact accuracy and coverage values are shown within bars. The classes (x axis) refer to the class as predicted by the model. NaN indicates that the model never predicted such a class

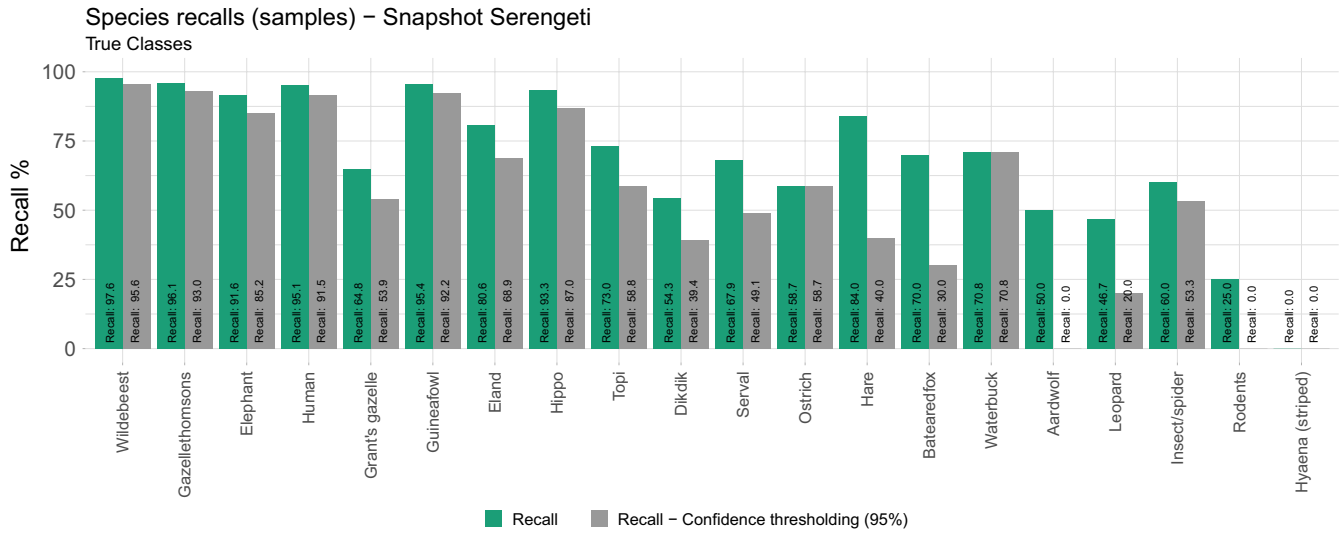

FIGURE 8 Species recalls for the SS species model (only a selection of species are shown)—with all capture events (green) and only high confidence events (grey, 95% confidence or higher). Exact recall values are shown within bars. The classes (*x* axis) refer to the true class as aggregated from annotations by citizen scientists

Figure 9 shows for each class (as predicted by the model) the proportion of capture events with expedited retirement, the model's precision for that class, and the recall of that class. Model precision was always numerically larger than expedited retirement because the latter was only possible if the model correctly classified a capture event (Conversely, it is possible that the model assigned an accurate species class, but expedited retirement was prevented by an erroneous volunteer classification). We note some significant differences between class precisions and class recalls: While "eland" classifications (a rare species) by the model were correct in 93.8% of all cases, only 17.2% of all "eland" capture events were detected by the model. This is an effect of confidence thresholding which led to a large portion of potential "eland" cases being ignored because the model was not certain enough, essentially trading recall for high precision. The model's precision for identifying empty images was 87.1% which is the only class with recall (87.9%) exceeding precision. This indicates that the model is more likely to miss an animal rather than detecting one that is not present.

4 | DISCUSSION

We investigated how deep learning can leverage manual classifications from citizen scientists by training powerful models which can

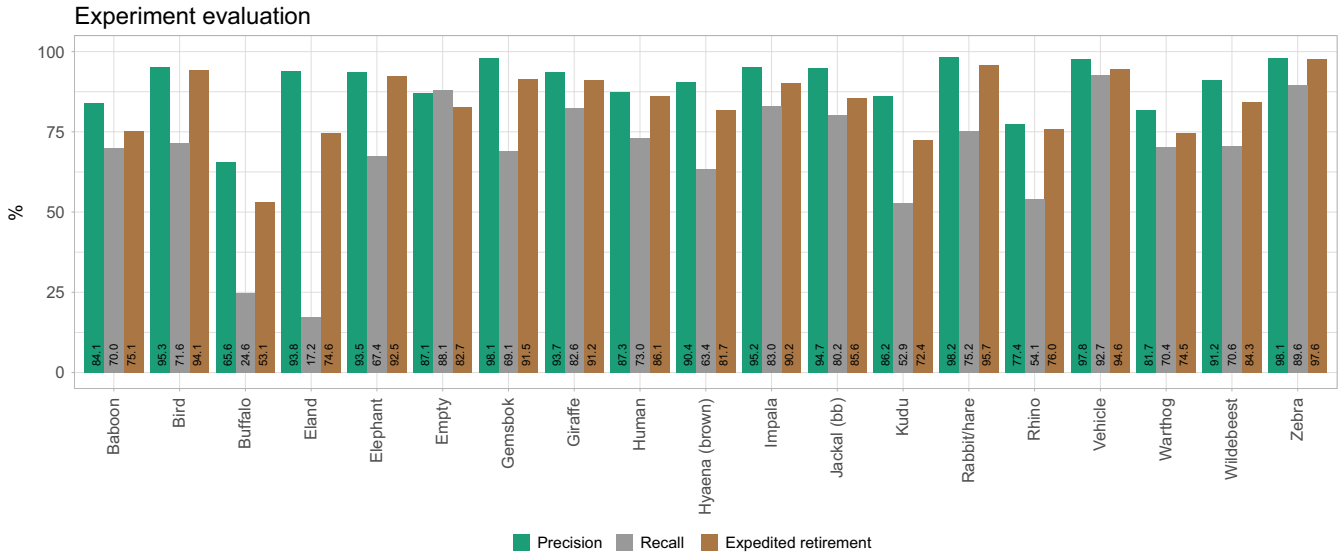

FIGURE 9 Shown are class-specific evaluations of the simulated experiment. Model precision (green bar) indicates the proportion of correct predictions per predicted class (in contrast to accuracy which is measured against the true class). Recall (grey bars) indicates the proportion of the true class correctly predicted by the model. Expedited retirement is indicated by the brown bars

help ecologists classify camera trap datasets. Our results show that CNNs reach high accuracies in camera trap classification on datasets labelled by citizen scientists. In particular, confidence thresholding improved model accuracy to the level of average human consensus (Swanson et al., 2016) for large segments of the datasets. Applying our models in a real-time setting has shown that 1–2 human annotations per image, if combined with the model prediction, can be enough to obtain accurate labels. Transfer-learning increased model accuracy substantially for small datasets and decreased model training time. Our results show that researchers can significantly decrease current processing times of camera trap datasets, thus paving the way to conduct very large studies while evaluating them in a timely manner.

4.1 | Model accuracies

Researchers can adjust confidence thresholds to control expected model accuracy by ignoring low-confidence predictions. Many of the most frequent classes reached near 100% accuracy when applying confidence thresholding without significant coverage reduction. Accuracy for rare species reached high values in some cases, however, coverage and recall were usually significantly reduced. Depending on the specific goals of a study and based on model performances evaluated on their test set it may be possible to rely solely on model predictions for final classification. If only specific species are of interest, the model could be used to identify possible candidates for manual confirmation by using low confidence thresholds to ensure high species coverage.

In most cases, transfer-learning surpassed the accuracy of models trained from scratch, even if species in the base model did not coincide with species in the target dataset. Furthermore, the benefit of using transfer-learning increased on smaller datasets, which implies that researchers can train and use a CNN much earlier in the labelling process because fewer annotated images are required. Models could be shared among different research teams and be quickly adapted to new datasets.

Our model accuracy on the SS dataset (92.5% for species identification) is similar to the results reported by Norouzzadeh et al. (2018) (90.4% on ResNet18). We surpassed the best results reported by Gomez Villa et al. (2017) (roughly 58% as estimated from their plots) who also used the SS dataset but excluded rare species (modelling 26 species) and used transfer-learning to train all of their models. However, contrary to our approach, they pretrained their models on the ImageNet dataset while we used transfer-learning between camera trap projects.

4.2 | Limitations

Classification accuracy was low for rare species, which reveals a clear limitation of CNNs. The expected accuracy in classifying a specific species, however, depends on multiple factors such as the absolute and relative occurrence of the species in the dataset, as well as the individual species (i.e., some species are easier to model than others). This makes it difficult to estimate how the number of annotated images translates into model accuracy. In general, the problem of low accuracies for rare species can be mitigated by considering prediction thresholds to identify cases with low model confidence and thereby ensure such cases can be reviewed by citizen scientists or experts. Norouzzadeh et al. (2018) addressed this problem by placing higher weights on rare classes during model training but were not able to systematically improve accuracies for rare species (see Buda, Maki, and Mazurowski (2018) for strategies to address class imbalance in modelling).

After training a model, its application on new data should be monitored carefully. One potential problem is that new types of images may be collected—images the model was not confronted with during model training—such as images with a new camera angle, new

species or from new locations. In such situations researchers should carefully assess model performance and re-train models if needed. Researchers should also be careful with equating the output of a CNN to real probabilities. Studies have shown that deep neural networks tend to be overly confident in their predictions (Guo, Pleiss, Sun, & Weinberger, 2017).

The aim of an ecology project may go beyond simple species identification. Researchers might want to additionally evaluate animal behaviour, differentiate between adults and young, male or female individuals, or obtain count estimates of animals in a particular image. Though some studies used models to localize and count different animals in images, they required datasets with bounding boxes or other localized annotations in order to train a model (Parham & Stewart, 2016; Zhang, He, Cao, & Cao, 2016). Norouzzadeh et al. (2018) demonstrated, however, that categorical (nonlocalized) labels can be sufficient to train a model to jointly identify species and other labels provided by citizen scientists, such as count or animal behaviour categories.

4.3 | Outlook

A next step for citizen science platforms like Zooniverse or research teams could be to automate camera trap classification. Integrating pretrained models into their data-processing pipelines and combining them with annotations from experts or citizen scientists as they become available (referred to as online-learning) could further reduce processing times. Ideally, such a process would produce probabilistic estimates of classifications to make better decisions in how to handle them, for example, whether to have them reviewed by an expert. Ideas on how to aggregate annotations from several, independent citizen scientists and combine them with model predictions is an active area of research (e.g., Branson, Horn, & Perona, 2017; Simpson, Roberts, Psorakis, & Smith, 2013). Ideally, the output of a model should represent probability distributions over all classes, which is an active area of research in Bayesian neural networks (e.g., Gal, 2016).

With increasing dataset sizes, computational capabilities, and improved model architectures, we expect accuracies of deep learning models to further increase. A promising field of research aims to reduce the number of training labels required, for example, Gal and Ghahramani (2015), which would be particularly beneficial for rare species. Furthermore, parameter tuning and model ensembling usually yield better results, for example, Norouzzadeh et al. (2018). Also, for projects with multiple images per capture event, motion patterns between subsequent images can contain valuable information to improve the detection of empty images (see Supporting Information for a case study) or to additionally classify animal behaviour.

Combining deep learning with the effort of citizen scientists may enable large camera trap studies to substantially expand their networks. For example the SS study operated over 200 camera traps in 2013 and required nine volunteer classifications per capture event to annotate those data. Using the approach described in this article it is feasible that capture events could be accurately labelled using as few as two volunteer classifications. This would enable data from 900 camera traps to be processed by the same cohort of citizen scientists within a similar time-frame. However, it is important to consider how volunteer experience changes when citizen science projects incorporate machine predictions and how that affects the project's success, quality, and number of annotations (see Bowyer, Maidel, Lintott, Swanson, and Miller (2015) for impact on volunteer engagement when controlling for the proportion of empty images). We note that the implementation and usage of such models is not a trivial task and therefore aim to provide the necessary software and guidelines to effectively assist ecologists who wish to apply them (see section 7 for more information). Our results validate the adoption of machine learning as a viable strategy to address the overwhelming data challenge facing ecologists and strongly motivate the requisite investment of time and effort. Moreover, we demonstrate the great potential of powerful machine learning algorithms to reinforce citizen scientist analysis of large camera trap studies.

ACKNOWLEDGEMENTS

We thank Hugh Dickinson, Chris Lintott, Sarah Pati, Laura Trouille, and Mike Walmsley for reviewing the manuscript. We also thank Jennifer Stenglein and other members of the SW team for program and data management. EE was funded by the University of Oxford's Hertford College Mortimer May fund. We thank ANPN Gabon, U. Stirling, J. Edzang-Ndong, D. Lehmann, Yadvinder Malhi, Imma Oliveras, William Bond, and Katharine Abernethy for contributing to EE. This study was partially supported by the NSF under award IIS 1619177. The development of the Zooniverse platform was partially supported by a Global Impact Award from Google. We also acknowledge support from STFC under grant ST /N003179/1.

AUTHORS' CONTRIBUTIONS

M.W. designed, implemented and evaluated the methods and led the writing of the manuscript; A.W.C, R.T.P, A.S., and C.L provided camera trap data; A.B and M.V. implemented software to run experiments; L.F. supervised the work. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

The code conducted to perform the study (https://doi.org/10.5281/zenodo.1426139), as well as curated code (https://doi.org/10.5281/zenodo.1434474) recommended for use by researchers were published via Zenodo. The models and additional meta-data (https://doi.org/10.13020/D6P67B), as well as most camera trap images (https://doi.org/10.13020/D6T11K) were published on the *Data Repository for the University of Minnesota (DRUM)*. Data for EE is not publicly available to protect endangered species from poaching but can be requested for research purposes via the *Oxford Research Archives (ORA)* (https://doi.org/10.5287/bodleian:mvv7jQnrR).

ORCID

Marco Willi http://orcid.org/0000-0002-0041-396X

Anabelle W. Cardoso http://orcid.org/0000-0002-4327-7259

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Brain, G. (2016). TensorFlow: A system for large-scale machine learning TensorFlow: A system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), pp. 265–284.

Bowyer, A., Maidel, V., Lintott, C., Swanson, A., & Miller, G. (2015). This image intentionally left blank: Mundane images increase citizen science participation. Retrieved from https://www.humancomputation.com/2015/papers/45_Paper.pdf

Branson, S., Horn, G. V., & Perona, P. (2017). Lean Crowdsourcing: Combining Humans and Machines in an Online System, Cvpr 7474–7483.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., Lintott, C., … Wallin, J. (2012). Galaxy zoo: Morphological classification and citizen science. In M. J. Way, J. D. Scargle, K. M. Ali, & A. N. Srivastava (Eds.), *Advances in machine learning and data mining for astronomy* (pp. 213–236). Boca Raton, FL: CRC Press.

Gal, Y. (2016). Uncertainty in Deep Learning, PhD Thesis, 174.

Gal, Y., & Ghahramani, Z. (2015). Bayesian convolutional neural networks with Bernoulli approximate variational inference. Retrieved from https://arxiv.org/pdf/1506.02158.pdf

Giraldo-Zuluaga, J. H., Salazar, A., Gomez, A., & Diaz-Pulido, A. (2017). Recognition of mammal genera on camera-trap images using multilayer robust principal component analysis and mixture neural networks. In *2017 IEEE 29th international conference on tools with artificial intelligence* (pp. 53–60). Piscataway, NJ: IEEE.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the 13th international conference on artificial intelligence and statistics (AISTATS)* (Vol. *9*, pp. 249–256). Proceedings of Machine Learning Research (PMLR).

Gomez Villa, A., Salazar, A., & Vargas, F. (2017). Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, *41*, 24–32. https://doi.org/10.1016/j.ecoinf.2017.07.004

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Retrieved from https://arxiv.org/pdf/1706.04599.pdf

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Piscataway, NJ: IEEE.

Hines, G., Swanson, A., Kosmala, M., & Lintott, C. (2015). Aggregating user input in ecology citizen science projects. In D. Gunning & P. Z. Yeh (Eds.), *Proceedings of the twenty-seventh conference on innovative applications of artificial intelligence* (pp. 3975–3980). Palo Alto, CA: The AAAI Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). La Jolla, CA: NIPS.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541–551. https://doi.org/10.1162/neco.1989.1.4.541

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*, E5716–E5725. https://doi.org/10.1073/pnas.1719367115

O'Connell, A. F., Nichols, J. D., & Karanth, K. U. (2011). *Camera traps in animal ecology: Methods and analyses*. Berlin, Germany: Springer Science & Business Media. https://doi.org/10.1007/978-4-431-99495-4

Parham, J., & Stewart, C. (2016). Detecting plains and Grevy's Zebras in the realworld. In *2016 IEEE winter applications of computer vision workshops (WACVW)* (pp. 1–9). Piscataway, NJ: IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution.*, *24*, 467–471. https://doi.org/10.1016/j.tree.2009.03.017

Simpson, E., Roberts, S., Psorakis, I., & Smith, A. (2013). Dynamic bayesian combination of multiple imperfect classifiers. In T. V. Guy, M. Karny, & D. H. Wolpert (Eds.), *Decision making and imperfection* (pp. 1–35). Berlin, Germany: Springer.

Swanson, A., Kosmala, M., Lintott, C., & Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology*, *30*, 520–531. https://doi.org/10.1111/cobi.12695

Swanson, A. A., Kosmala, M., Lintott, C. C., Simpson, R. 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*, 150026. https://doi.org/10.1038/sdata.2015.26

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 3320–3328). La Jolla, CA: Neural Information Processing Systems (NIPS).

Yu, X., Wang, J., Kays, R., Jansen, P. A., Wang, T., & Huang, T. (2013). Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, *2013*, 52. https://doi.org/10.1186/1687-5281-2013-52

Zevin, M., Coughlin, S., Bahaadini, S., Besler, E., Rohani, N., Allen, S., … Kalogera, V. (2016). Gravity spy: Integrating advanced LIGO detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, *34*, 1–27.

Zhang, Z., He, Z., Cao, G., & Cao, W. (2016). Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. *IEEE Transactions on Multimedia*, *18*, 2079–2092. https://doi.org/10.1109/TMM.2016.2594138

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Willi M, Pitman RT, Cardoso AW, et al. Identifying animal species in camera trap images using deep learning and citizen science. *Methods Ecol Evol.* 2019;10:80–91. https://doi.org/10.1111/2041-210X.13099