CITIZEN SCIENCE:
THEORY AND PRACTICE

Human Supervision is Key to Achieving Accurate AI-assisted Wildlife Identifications in Camera Trap Images

COLLECTION: THE FUTURE OF ARTIFICIAL INTELLIGENCE AND CITIZEN SCIENCE

METHOD

]u[ubiquity press

SARAH E. HUEBNER
MEREDITH S. PALMER
CRAIG PACKER

*Author affiliations can be found in the back matter of this article

ABSTRACT

Using public support to extract information from vast datasets has become a popular method for accurately labeling wildlife data in camera trap (CT) images. However, the increasing demand for volunteer effort lengthens the time interval between data collection and our ability to draw ecological inferences or perform data-driven conservation actions. Artificial intelligence (AI) approaches are currently highly effective for species detection (i.e., whether an image contains animals or not) and labeling common species; however, it performs poorly on species rarely captured in images and those that are highly visually similar to one another. To capitalize on the best of human and AI classifying methods, we developed an integrated CT data pipeline in which AI provides an initial pass on labeling images, but is supervised and validated by humans (i.e., a "human-in-the-loop" approach). To assess classification accuracy gains, we compare the precision of species labels produced by AI and HITL protocols to a "gold standard" (GS) dataset annotated by wildlife experts. The accuracy of the AI method was species-dependent and positively correlated with the number of training images. The combined efforts of HITL led to error rates of less than 10% for 73% of the dataset and lowered the error rates for an additional 23%. For two visually similar species, human input resulted in higher error rates than AI. While integrating humans in the loop increases classification times relative to AI alone, the gains in accuracy suggest that this method is highly valuable for high-volume CT surveys.

CORRESPONDING AUTHOR:
Sarah E. Huebner
Smithsonian Conservation Biology Institute, US
huebners2@si.edu

KEYWORDS:
ecology; conservation; camera traps; Snapshot Safari; human-in-the-loop; artificial intelligence

TO CITE THIS ARTICLE:
Huebner, SE, Palmer, MS and Packer, C. 2024. Human Supervision is Key to Achieving Accurate AI-assisted Wildlife Identifications in Camera Trap Images. *Citizen Science: Theory and Practice,* 9(1): 38, pp. 1–12. DOI: https://doi.org/10.5334/cstp.752

INTRODUCTION

The cumulative impacts of human activities are creating rapid changes in the natural world, necessitating frequent measurements and updates to conservation management strategies (Pacifici et al. 2015; Bonebrake et al. 2018; Schlaepfer and Lawler 2023). To understand and mitigate these challenges on effective timescales, ecologists and conservation practitioners require timely access to large amounts of data on wildlife populations. Remote sensors are increasingly deployed to evaluate how habitats and species respond to global or local changes and management or conservation interventions (Anderson et al., 2016; Davis et al., 2023; Kays et al., 2010; Palmer et al., 2022). Among these sensors, motion-activated wildlife cameras or "camera traps" are widely used due to their ability to collect detailed data on animal presence, behavior, health, and population structure with minimal disturbance. Since 2010, the number of ecologists using camera traps has grown exponentially due to the hardware's relatively low cost, ease of use, and commercial availability, accelerating the pace and volume of ecological data collection and aggregation (Burton et al. 2015; Steenweg et al. 2017). This work has laid the foundations for novel modeling approaches and large-scale inference; however, the power and accuracy of these insights rely on the classification accuracy of raw camera trap (CT) image data (Burton et al. 2015; Caravaggi et al. 2020; Farley et al. 2018; Hofmeester et al. 2020; Wevers et al. 2021; Laporte-Devylder et al. 2023).

Surveys composed of dozens to hundreds of CTs may produce tens of thousands of images over several months, which must subsequently be labeled with the observed species, counts of individual animals, and other variables pertinent to the research or conservation action in question. In the gathered datasets, many "empty" images triggered, for example, by windblown plants or overheated electronics, must be labeled as such and set aside. Even the resulting "animals only" dataset may contain numerous species that are not the focal species relevant to the research or conservation question at hand. This scenario is becoming commonplace as long-term ecological monitoring is adopted to track and combat the current extinction crisis (Ceballos et al. 2017).

Managing such volumes of data can rapidly exceed the capacity of most CT practitioners, leading many to request volunteer help (i.e., citizen or participatory science) via online platforms, such as Zooniverse.org, that facilitate public participation in scientific research (Gadsden et al., 2021; Hsing et al., 2018; Islam and Valles, 2020; Jones et al., 2018). The Zooniverse is the world's largest citizen science platform, with greater than 2.7 million registered users who assist projects across multiple scientific disciplines, including ecology, astronomy, and biomedical research (Simpson et al., 2014; Trouille et al., 2019). The first wildlife image–based project to launch on Zooniverse was Snapshot Serengeti, a survey of 225 CTs running continuously since 2010 in Serengeti National Park, Tanzania (Swanson et al. 2015; Palmer et al. 2021). Using a suite of tools to narrow the potential species options by body shape, horns, pelage, color, and patterns, and other distinguishing characteristics, citizen scientists were asked to categorize the species observable in each picture along with the number of individuals and basic behaviors and demographics. Separate details were collected about every species in an image when multiple species were visible. Each image was classified by up to 20 volunteers whose responses were aggregated. The species receiving the most votes was accepted as the final label.

When the public-facing webpage launched in 2013, the project had a backlog of 1.2 million observations that were processed by volunteers in three days (Swanson et al., 2015). In the subsequent decade, more than 120 similar CT projects have joined Zooniverse, including 24 additional project pages for the expanded Snapshot Safari network launched in 2018, in which Serengeti is the flagship survey (Pardo et al. 2021). Snapshot Safari is the largest use case on Zooniverse, with 50+ CT surveys in six African countries collecting data using the same protocols. (Surveys with < 10 CTs and data from temporary CT sites are folded into fewer project pages on Zooniverse to facilitate data processing pre- and post-classification.) It is also popular with volunteers, with > 200,000 users from 77 countries logging on regularly to assist with wildlife image classification. Still, increased competition for a finite number of willing volunteers extends the interval necessary to obtain fully and accurately labeled image datasets and return the information they contain to researchers and conservation managers.

Increasingly, improvements in computational power and the lengthening time required to obtain image labels through citizen science have motivated researchers to develop artificial intelligence (AI) models for quicker data classification (Norouzzadeh et al. 2017; Borowiec et al. 2022; Vélez et al. 2023). This presents its own set of challenges, as training a highly accurate species-classifying algorithm (hereafter, "classifier") has traditionally demanded large volumes of labeled data, on the order of thousands of images per species category ("class"). This process therefore requires millions of labeled images (i.e., a "many shot" approach) for highly diverse ecosystems like the African savanna (Pantazis et al., 2024). A typical CT survey produces data with a long-tailed distribution, in which a few classes contain the vast majority of images while most classes contain relatively few (Miao et al. 2021). Thus, AI models often perform better at identifying highly abundant or visible species than rare and cryptic species that are captured less frequently or threatened species with low population sizes.

METHODS

The Snapshot Safari network first deployed AI in 2018, using labeled image data from the Serengeti CT grid and eight other African CT surveys to create a custom classifier for processing the high volumes of data produced by continuous long-term monitoring (Willi et al., 2019). These models are available on GitHub for open access use (See Data Availability Statement 1). Due to Snapshot Serengeti's early entry into the field, many existing classifiers for African wildlife (including transfer learning techniques and broader extensions of CT data) have been built using the training data labeled by volunteers (Battu and Reddy Lakshmi, 2023; Liu et al., 2024; Norouzzadeh et al., 2021; Villa et al., 2017). All publicly available data is posted to the Labeled Image Library of Alexandria for Biology and Conservation (see Data Availability Statement 2). Although rigorously assessed in an academic context, the specialized Snapshot Safari classifier was largely untested in practice. As such, we opted for a "humans-in-the-loop" (HITL) approach when integrating the AI model into our pipeline to ensure the accuracy of labels prior to use for research and conservation purposes.

Prior to incorporating AI, all image labels were generated by Zooniverse citizen scientists. Capture events consisting of one to three images (three taken in rapid succession during the day and one at night) were evaluated by up to 20 volunteers before retirement (removal from circulation), and a consensus algorithm was run to determine the species with the highest number of votes, which became the ultimate label (Swanson et al. 2015). All images within a capture event were kept together throughout the classification process as they constitute a single observation, and more data may be gleaned from a series of photos than one.

Though the urgent necessity of returning processed data quickly has made it no longer possible to rely on citizen science alone, AI-generated labels are known to be fallible, and many researchers recommend carefully evaluating these predictions to assess accuracy across time, space, and species (Clarfeld et al., 2023; Whytock et al., 2021). AI accuracy varies for many reasons, including low numbers of training images, differing background vegetation, difficult-to-interpret images (Westworth et al., 2022), and/or model drift, that is, the tendency of algorithms to become outdated and lose accuracy over time, which may necessitate additional transfer learning and training (Ackerman et al., 2021). Prior work has documented that volunteers are accurate 96.6% of the time across Serengeti species classes (Swanson et al., 2016), demonstrating that humans effectively identify species even when they have never or rarely viewed them. Therefore, we developed a hybrid approach to classifying images in which human volunteers supervised the classifier's labels, that is, a HITL strategy. Here, we compare labels produced by AI and HITL against a newly created gold standard (GS) dataset annotated by professional ecologists to evaluate the impact of human guidance on final precision and accuracy.

GOLD STANDARD DATA

To evaluate the accuracy of both classification methods, experienced wildlife researchers (+3 years of experience identifying African animals) classified the species in captures sampled from data collected from early 2015 to mid 2017. This is an update to the original Snapshot Serengeti GS dataset published in 2016 (see Data Availability Statement 3). We selected 6,000 capture events across species classes by including all available records of rare species during that timeframe and evenly distributing the remainder across commonly observed species, resulting in 44 species classes. When the expert classified a capture as "unresolvable" (cannot be classified because the image is too dark, there is too much motion blur, an animal was occluded by vegetation, etc.), we removed it from this analysis since it is impossible that either AI or HITL would converge on the correct answer if the expert could not. This excluded 216 captures or 3.6% of the newly created GS dataset. We also excluded 470 empty captures (7.83% of GS dataset) from this analysis because we are focused on the accuracy of species labels as the most important contribution of human supervision. This resulted in 5,314 species comparison points from the expert labels.

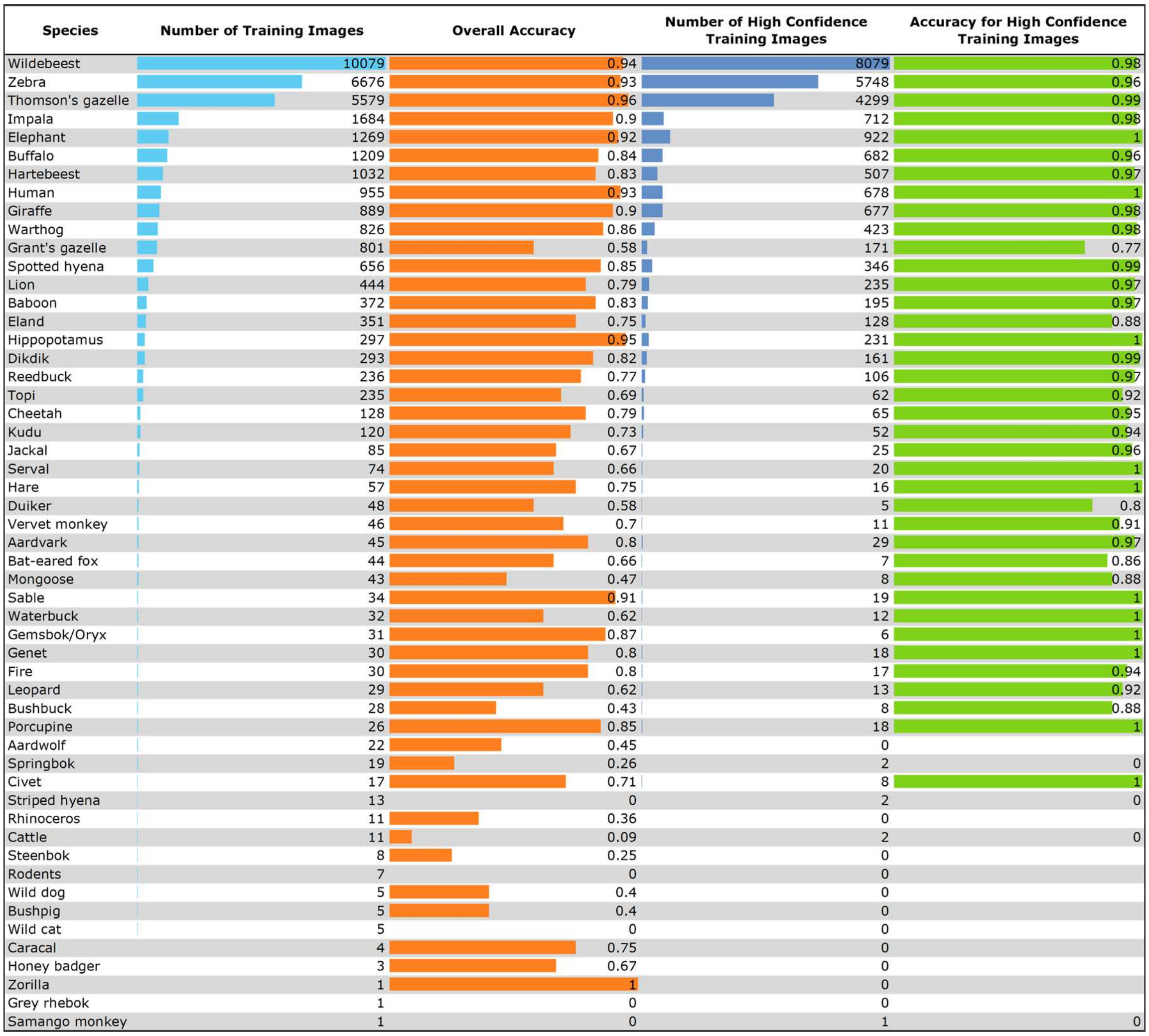
SNAPSHOT SAFARI AI MODELS

The Snapshot Safari data pipeline incorporates labels generated by two convolutional neural network (CNN) models trained using the ResNet-18 model architecture (He et al., 2016) and Tensorflow (Abadi et al., 2016). Training data comprised 3.66 million CT images from nine African protected areas participating in Snapshot Safari (five in South Africa, three in Tanzania, and one in Mozambique). CNN models learn the intrinsic features of training data by updating model parameter weights and subsequently outputting inference results or "predictions" on unseen data; therefore, training data distribution has a crucial impact on model performance (Pantazis et al., 2024). Serengeti data constituted 89% of the training images because these models were created shortly after the inception of Snapshot Safari and contained only the first six months of data from the other CT grids. Therefore, we focused on the Serengeti GS dataset for these analyses. The classifier was trained by randomly assigning all images to one of three datasets while preserving class distributions: training (90% of the data), validation (5%), and testing (5%). The models learned from the training set, and were monitored on the validation set to reduce overfitting (Willi et al., 2019).

The Snapshot Safari object detection CNN ("detector") was trained to flag empty images, which can constitute a significant percentage of images from CTs situated in grasslands or other biomes with rapid vegetation growth. It produces a binary output of "empty" or "not empty" and a confidence (probability) score indicating certainty in its prediction. The detector was evaluated on 71,702 test samples, of which 74.38% were rated as "high confidence," set at a minimum threshold of 95%. The detector achieved an overall accuracy score of 96.03% rising to 99.5% on high-confidence images.

The species-classifying CNN ("classifier") generates predictions with associated confidence scores (the probability of a match for every species class based on the classifier's perception of similarity) for 56 species classes. The classifier was assessed on 36,469 test captures, 69.53% of which were labeled as high confidence, again set at 95% minimum threshold. As differences in regional wildlife morphology, topographic heterogeneity, and background vegetation impact AI accuracy (Beery et al., 2018), we sampled a diverse mixture of training photos across sites, seasons, and species while selecting all available images of rarely seen species. The species assemblages are similar at most of the Snapshot Safari sites in eastern and southern Africa, but regional variations in morphology and colloquial species names exist, so we harmonized the taxonomies prior to training.

The overall model accuracy was 89.4%, rising to 97.2% for high-confidence images. The three most commonly observed species in the Serengeti—wildebeest (*Connochaetes taurinus*), plains zebra (*Equus quagga*), and Thomson's gazelle (*Eudorcas thomsonii*)—accounted for ~65% of Serengeti training images (see Figure 1). Wildebeest images were 94% accurate across the board, which improved to 98% accuracy on images rated as high



Species	Number of Training Images	Overall Accuracy	Number of High Confidence Training Images	Accuracy for High Confidence Training Images
Wildebeest	10079	0.94	8079	0.98
Zebra	6676	0.93	5748	0.96
Thomson's gazelle	5579	0.96	4299	0.99
Impala	1684	0.9	712	0.98
Elephant	1269	0.92	922	1
Buffalo	1209	0.84	682	0.96
Hartebeest	1032	0.83	507	0.97
Human	955	0.93	678	1
Giraffe	889	0.9	677	0.98
Warthog	826	0.86	423	0.98
Grant's gazelle	801	0.58	171	0.77
Spotted hyena	656	0.85	346	0.99
Lion	444	0.79	235	0.97
Baboon	372	0.83	195	0.97
Eland	351	0.75	128	0.88
Hippopotamus	297	0.95	231	1
Dikdik	293	0.82	161	0.99
Reedbuck	236	0.77	106	0.97
Topi	235	0.69	62	0.92
Cheetah	128	0.79	65	0.95
Kudu	120	0.73	52	0.94
Jackal	85	0.67	25	0.96
Serval	74	0.66	20	1
Hare	57	0.75	16	1
Duiker	48	0.58	5	0.8
Vervet monkey	46	0.7	11	0.91
Aardvark	45	0.8	29	0.97
Bat-eared fox	44	0.66	7	0.86
Mongoose	43	0.47	8	0.88
Sable	34	0.91	19	1
Waterbuck	32	0.62	12	1
Gemsbok/Oryx	31	0.87	6	1
Genet	30	0.8	18	1
Fire	30	0.8	17	0.94
Leopard	29	0.62	13	0.92
Bushbuck	28	0.43	8	0.88
Porcupine	26	0.85	18	1
Aardwolf	22	0.45	0	
Springbok	19	0.26	2	0
Civet	17	0.71	8	1
Striped hyena	13	0	2	0
Rhinoceros	11	0.36	0	
Cattle	11	0.09	2	0
Steenbok	8	0.25	0	
Rodents	7	0	0	
Wild dog	5	0.4	0	
Bushpig	5	0.4	0	
Wild cat	5	0	0	
Caracal	4	0.75	0	
Honey badger	3	0.67	0	
Zorilla	1	1	0	
Grey rhebok	1	0	0	
Samango monkey	1	0	1	0

Figure 1 Evaluation of Snapshot Safari classifier performance on images in the test dataset for every possible species class from the nine sites that provided training data. The model was evaluated for accuracy across all training images and separately for images classified with high confidence scores (>95%).

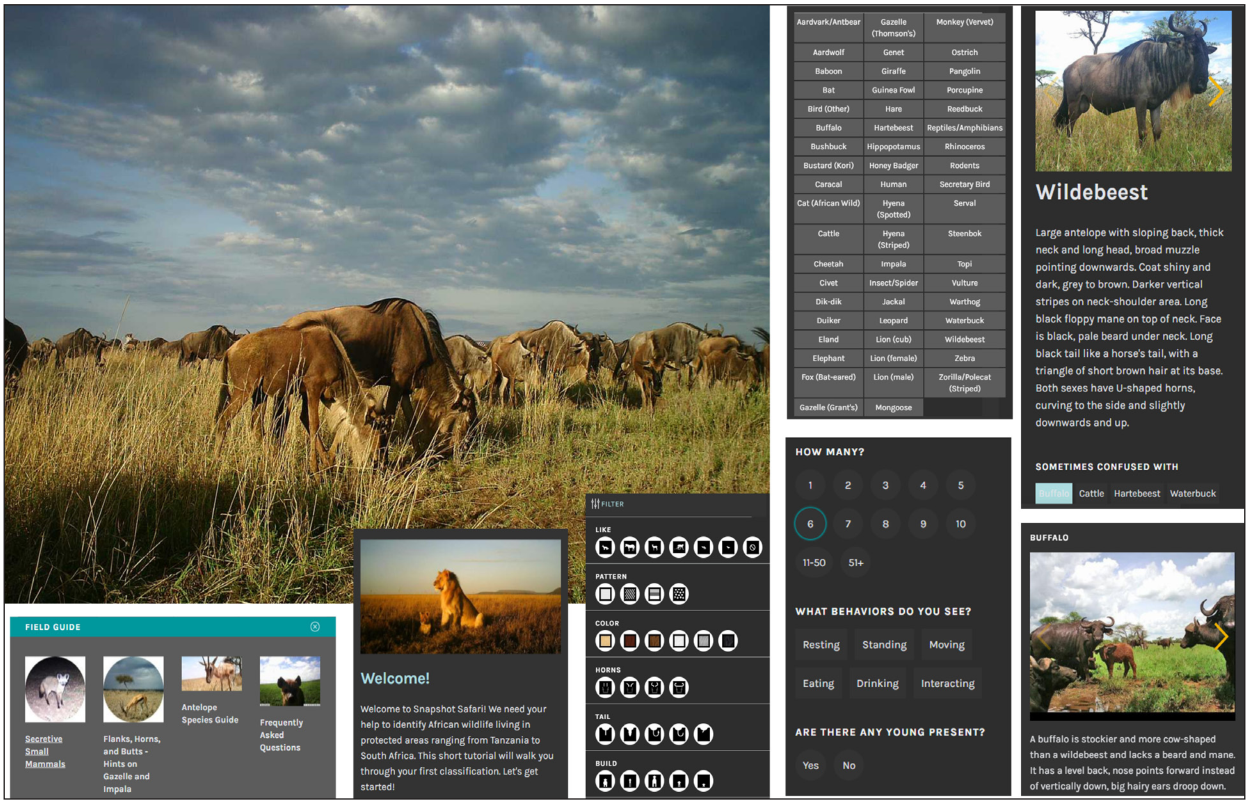
confidence. For the wildebeest class, 80% of the images in the testing dataset were marked as high confidence by the classifier. Zebra were correctly identified in 93% of all images and 96% of high-confidence images (86% of test set), and Thomson's gazelles were correct 96% of the time, improving to 99% on high-confidence images (77% of test set). In contrast, 15 species classes returned fewer than five available training images per class, resulting in 0% accuracy for 11 classes in the testing stage. Two species within this group returned 40% accuracy—bushpigs (*Potamochoerus larvatus*) and wild dogs (*Lycaon pictus*)—but zero images were classified as high confidence.

With these species-specific accuracies in mind, we determined that species classes in which many subjects are returned with high confidence scores can be processed more quickly using human supervision as a guardrail to ensure accuracy, whereas low-confidence images should be reviewed by more people when consensus is not achieved early in the classification process. Images of species for which the classifier demonstrated low accuracies were circulated for additional votes to ensure human volunteers converged on the correct label, which also helped to generate additional training data.

SNAPSHOT SAFARI HITL PIPELINE

To combine our citizen science and AI processing pipelines, we presented volunteers with images that had also received AI classifications. (See Palmer et al. (2021) for a more detailed description and flow chart of the pipeline.) Volunteers then completed two stages of labeling—detection and classification. The detection stage consisted of a question task, in which volunteers were asked to evaluate whether animals were present in the capture ("Empty or not?"). Where once as many as 20 people needed to vote on whether a capture was empty to ensure an accurate label, adding AI predictions allowed for the reduction of volunteer effort at this stage, provided they agreed with the detector's label. If two people agreed with the AI prediction that a capture was empty, it was labeled empty. If one of the first two people to view a capture disagreed with the AI label, it remained in circulation to accumulate three more human votes, at which point the majority consensus was accepted.

Captures marked as containing wildlife were moved to the classification phase, in which species identity was annotated. Within the species classification menu, volunteers compared a CT image with several exemplar



Large antelope with sloping back, thick neck and long head, broad muzzle pointing downwards. Coat shiny and dark, grey to brown. Darker vertical stripes on neck-shoulder area. Long black floppy mane on top of neck. Face is black, pale beard under neck. Long black tail like a horse's tail, with a triangle of short brown hair at its base. Both sexes have U-shaped horns, curving to the side and slightly downwards and up.

A buffalo is stockier and more cow-shaped than a wildebeest and lacks a beard and mane. It has a level back, nose points forward instead of vertically down, big hairy ears droop down.

Welcome to Snapshot Safari! We need your help to identify African wildlife living in protected areas ranging from Tanzania to South Africa. This short tutorial will walk you through your first classification. Let's get started!

Figure 2 Exemplar images, descriptions, and comparison tools that are available to volunteers on the Snapshot Safari survey menu. Selecting Wildebeest in the species menu brings up the exemplar images and description of the species. If the volunteer is unsure about a classification, they can click on the species in the "Sometimes confused with" category, which pops up that species' images and description for comparison. Volunteers can also use field guides, tutorials, and filters to assign a label. They are then asked to count the individuals and annotate behaviors and basic demographics, including counting horns in dimorphic species.

images of each species from different angles, read detailed descriptions of species traits, and reviewed "confused with" pairs side by side to compare the likeness of two species, as depicted in Figure 2. They could also filter species by features such as body type and horn shape, and turn to detailed field guides that provided additional tips on selecting the correct species.

At this stage, more complex and dynamic rules were used to label captures based on the AI species label, the AI confidence score, the number of volunteer votes, and the agreement of the volunteers with the classifier and one another. The objective was to attain the correct species label as quickly as possible without sacrificing accuracy. For instance, if the classifier returned an 80% confidence score that an image contained a wildebeest (a frequently observed animal on which AI accuracy is typically high), it was assigned that label when the first five volunteers confirmed that a wildebeest was present. This rule was also implemented for zebras and impala (*Aepyceros melampus*), an antelope species with high abundances across Snapshot Safari field sites. Consensus and classification counts were dynamically checked as classification advanced. For most of the species classes, captures were retired as they reached a consensus of at least 50% at ten votes or 25% at 15 votes. If consensus >50% was not reached after 15 votes, that capture was reviewed by the research team to assign a label. Many images that failed to reach consensus captured an incomplete animal or were the result of the animal being too close to the camera. The three most common species were assigned labels after ten votes maximum. This reduction in the number of votes helped to save time and volunteer effort across the classification process. The ultimate label is the combined effort of AI and humans, described as HITL in the following analyses. The AI values are the top predictions produced by the classifier on unseen CT observations without further training or human assistance.

ANALYSIS AND RESULTS

Across all species classes, AI returned an error rate of 34.89%, and HITL reduced the error rate to 8.73%. As shown in Figure 3, the HITL method outperformed AI in 95% of mammal and large bird classes (42 of 44), resulting in a total error reduction of 26.15%. AI achieved its best performance on highly abundant species within the Serengeti, that is, classes for which many training examples were provided. The classifier was trained on 10,079 images of wildebeest and returned an initial error rate of 7.63% on newly presented CT data. HITL reduced the error rate to 5.93% when compared with the GS responses. There were 6,676 training images of zebra, for which AI produced an error rate of 5.63% on unseen images. Adding the volunteers' input reduced the error rate to 1.25%. Thomson's gazelles (5,579 training images) were a notable exception to this trend—AI produced an error rate of 7.24%, which rose to 9.5% in our evaluation of the HITL classification method. The largest improvements in error rates came in species classes that are easily recognizable to humans, such as

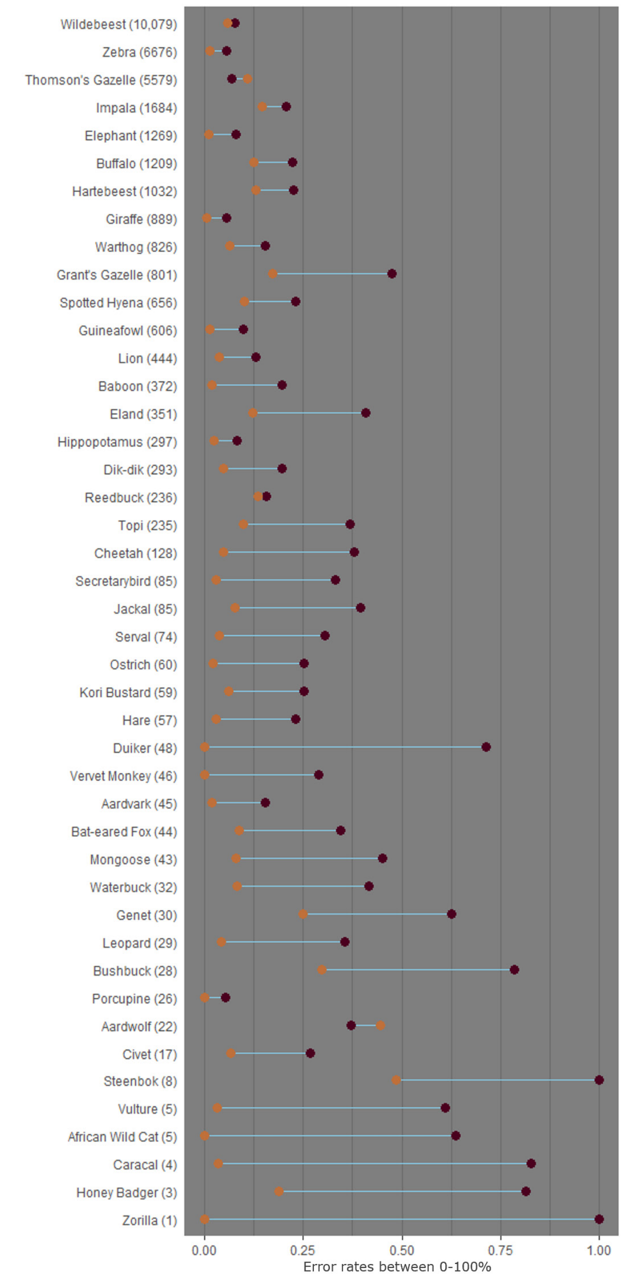


Figure 3 Differences in species-specific error rates between AI (*black circles*) and HITL (*orange circles*) classification methods. The blue bar represents the size of the error reduction from adding HITL to supervise AI classifications of unseen CT images. In all except two cases (aardwolf and Thomson's gazelle), HITL decreased the error rate. Species are arrayed from most to least common within the training dataset with the number of training images in parentheses.

giraffe (*Giraffa giraffa*), which improved from a 5.59% error rate in AI to 0.62% in HITL; hippopotamus (*Hippopotamus amphibius*), which improved from 8.24% to 2.35%; and cheetahs (*Acinonyx jubatus*), which improved from 37.95% to 4.82%.

For species for which fewer than 100 images were available to train the classifier, the HITL classification method yielded an average error rate reduction of 37.94% on novel CT data, except for aardwolves (*Proteles cristata*), in which the AI error rate of 37.19% was worsened to 44.63% by human classifications. For many of the species in this < 100 category, HITL improved final labels but still returned unacceptable levels of accuracy. For instance, images of steenbok (*Raphcerus campestris*), a small antelope that is rare in the Serengeti, returned a 100% error rate by AI and 48.57% by HITL. However, there were significant gains in accuracy in other rarely photographed species like leopards (*Panthera pardus*), which saw an error rate reduction from 35.48% for AI to 4.30% for HITL; ostriches (*Struthio camelus*) dropped from 25.17% to 2.10%; and waterbuck (*Kobus ellipsiprymnus*) decreased from 41.67% to 8.33%.

DISCUSSION

A species classifier confronted with a seldom-seen species, camera angle, or location is highly likely to make a classification error (Beery et al., 2018). Snapshot Safari runs a custom classifier trained on data generated by long-standing CT survey grids, yet HITL is still required to ensure accuracy for most species classes when AI is applied to new data from the same sites. The amount of training data is crucial since classifiers cannot identify animals without a plurality of examples to pull from. Humans, on the other hand, can use exemplar images, descriptions (including how to discern the differences between species that are morphologically similar), field guides, and assistance from researchers and moderators to make determinations.

The largest performance improvements achieved by HITL were in classes for which few training images were available, for example, predators and other cryptic species. With just one image of a zorilla (*Ictonyx striatus*) in the comprehensive training dataset, it is expected that AI will not be sufficiently effective for this class or others with similarly low availability of training images. The effect of small class size may be mitigated for species exhibiting unique morphological traits as in the case of crested porcupines (*Hystrix cristata*) in this dataset. With only 26 training images, the classifier returned an error rate of 5.15%, the lowest of any class. HITL reduced the error rate to zero.

Within the predator guild, with 29 capture events marked as “caracal” (*Caracal caracal*) by experts and just four caracal images in the training dataset, AI predicted the wrong species 82.76% of the time, whereas HITL reduced the error rate to 3.45%. Of the 24 errors by AI in caracal images, the most common confusion was lion (*Panthera leo*). Several classes that were confused by AI and HITL are shown in Figure 4. Caracals and lions are both medium-large cats with tawny coats, but humans readily distinguished between them even at night due to morphological differences in their ears and tails.

It is important to note that improvement from involving humans in the classification effort is context-dependent, and there is also an element of bias in volunteer responses. In two species classes evaluated here, humans performed worse than the classifier. In both cases, species with similar morphologies confused volunteers, particularly in images where only a portion of an animal is visible. HITL accuracy is aided by batch aggregation of classifications to obtain the label with the highest consensus, so correct responses normally outweigh one incorrect vote, but multiple wrong votes can lead to an incorrect label. The volunteers’ level of agreement is tracked in our reporting to each Snapshot Safari site to provide information to the project team about which captures achieve consensus and which may warrant further review. Anything less than 60% consensus among human classifiers likely needs to be checked before use in research studies. As demonstrated here, species with known confusions should also be reviewed by researchers because occasionally humans achieve consensus on the wrong identification.

One instance in which consensus did not reliably produce the correct label was the species category “aardwolf,” in which so many people selected “striped hyena” that the HITL error rate was higher than AI despite only 13 images in the training dataset (see Figure 4). This may be attributed to observer bias from volunteers who are eager to see striped hyenas and hence overestimate their perceived occurrence in the study area. Since many captures of both species occur at night and may not have a view of the tail, these images can be tough for experts to decipher, as well.

In the other instance, Thomson’s gazelle was the third most common species in the AI training dataset and returned an error rate of 7.24% in AI that rose to 9.5% via HITL. Thomson’s gazelles are similar in appearance to Grant’s gazelles (*Nanger granti*) and impala, which are sympatric throughout the species’ range. Most of the HITL errors for these three species reflected confusion among them (see Figure 4). It is unclear why Thomson’s gazelle, in particular, elicited so many incorrect responses despite extensive field guides on how to tell these three species apart.

AI accuracy can vary widely depending on the ecosystem, species, and number of images available for training models. If researchers are focused only on the

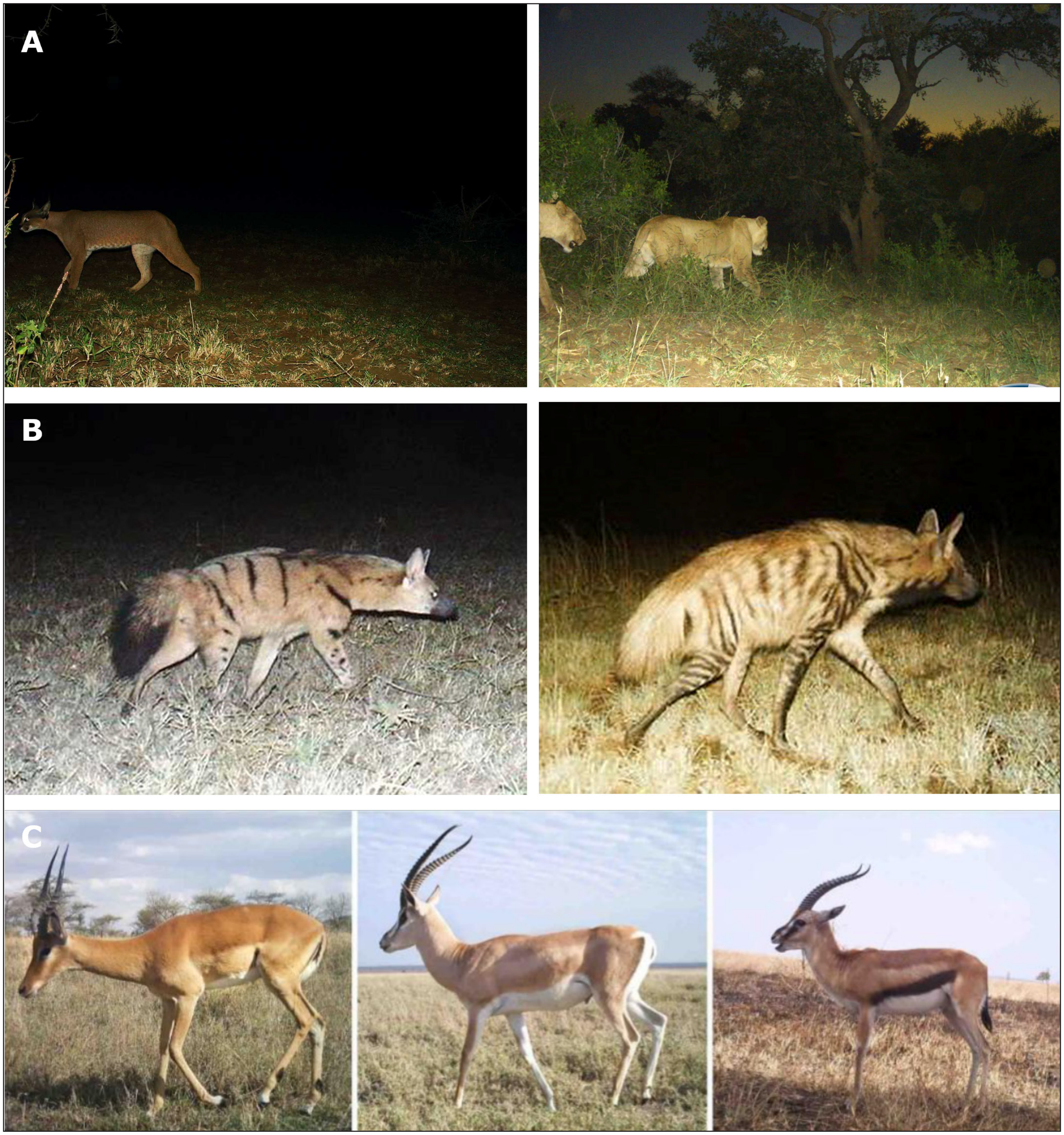


Figure 4 Comparison images of species commonly confused by AI (a) and HITL (b, c). **a)** AI consistently labeled pictures of caracals (*left*) as lions (*right*). **b)** HITL frequently labeled images of aardwolves (*left*) as striped hyenas (*right*). **c)** HITL reduced error rates for impala (*left*) and Grant's gazelle (center) but performed worse than AI on Thomson's gazelle (*right*).

most abundant species, AI trained on many images may be sufficient to meet their needs. For projects evaluating community dynamics or a suite of species, quickly removing empty images and the most abundant species classes from circulation allows people to spend more time reviewing images of wildlife that the classifier does not recall well.

With the proliferation of CT studies and field surveys using multitudes of ecological sensors, researchers are searching for ways to quickly and accurately translate raw data into usable data points. AI is an appealing solution because it allows for rapid translation once a model has been trained. It has until recently been a daunting task to build labeled image libraries; however, citizen science provides opportunities for labeling data using crowdsourcing through citizen science platforms. Additionally, recent machine-learning techniques have successfully developed models using fewer training images for select species classes (Schneider et al., 2020; Shahinfar et al., 2020). Generalized models that are agnostic to locations and backgrounds, such as the object detection model MegaDetector (Beery et al., 2019); Beery et al. 2019), and the global species classifier created by Wildlife Insights (Ahumada and Fegraus 2019) represent another encouraging trend.

Yet even these improved AI models require ongoing human training, validation, and supervision to return sufficiently accurate results to inform research studies and conservation management interventions. Ecologists using AI should ensure that model labels are verified and corrected by humans as necessary after initial evaluations of error rates by class. Our results indicate that this validation can be turned over to citizen scientists in most cases, but careful assessment of model and volunteer performance on a case-by-case basis is warranted. It is also important that researchers communicate with volunteers about the introduction of AI into data pipelines using citizen science. Some volunteers may be concerned that they are being replaced and their skills are no longer needed, which could lead to fewer engagements and shorter classification sessions. They should be informed that they are reviewing and correcting AI labels so that they do not become complacent. Further, communicating with volunteers about next-generation projects with which they can engage is crucial to retention of a motivated volunteer workforce. Experienced Snapshot Safari citizen scientists are now asked to take on tasks that help with identifying individual animals within populations or contributing behavioral labels from CT projects that had not previously extracted that information from existing image collections.

Citizen science and AI offer researchers the ability to quickly and efficiently move images through data pipelines to achieve labels and create data points for use in conservation programs and research studies. Citizen science can be a time-consuming process and requires monitoring of talk forums while projects are active. When AI was introduced to the Serengeti pipeline, classification time was halved even using conservative rules for getting enough volunteer votes to ensure accuracy. Times could be reduced further still by lowering confidence thresholds and the number of votes required to achieve a label. In species classes in which AI error rates are low, these images may not need to be presented to humans for help. The involvement of well-trained volunteers can greatly improve the performance of AI alone and is recommended, particularly for CT surveys with large volumes of data.

CONCLUSIONS

- HITL produced fewer errors than AI for 95% of species classes evaluated with the Serengeti GS dataset and decreased the overall error rate from 34.89% to 8.73%, with larger gains in species for which fewer than 100 training images were available.
- AI accuracy varied widely and was correlated with the amount of training data available for a particular species. Species classes with high volumes of images and a high proportion of high-confidence scores may be sufficiently accurate to warrant removing HITL for the most common classes in a long-tailed distribution in addition to removing empty images.
- Fewer training images result in worse outcomes from both methods, but humans are more likely to converge on the correct answers due to context clues and consensus.
- Turning over the HITL piece to volunteers can save researchers time while still yielding high-quality labeled image data.
- Researchers incorporating AI into pipelines with citizen science should communicate with volunteers about how their efforts benefit the project and about new tasks suited to humans that will lead to more informative data for biodiversity monitoring.

DATA ACCESSIBILITY STATEMENT

[1] Snapshot Safari trained CNN models:
The object detector and species classifier used for this work are available on GitHub: https://github.com/marco-willi/camera-trap-classifier.

[2] Labeled camera trap data:
Labeled Image Library of Alexandria – Biology and Conservation (LILA-BC) repository contains camera trap datasets from Snapshot Safari camera traps including data collected in Serengeti National Park (Tanzania), Camdeboo National Park, Karoo National Park, Kgalagadi Transfrontier Park, Kruger National Park, and Mountain Zebra National Park (South Africa) and the Enonkishu Conservancy (Kenya). New data is uploaded as classifications are provided by citizen scientists (maintained by Google AI): http://lila.science/datasets.

[3] Previous gold standard dataset for Snapshot Serengeti:
Swanson, Alexandra B. et al. (2016). Data from: Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna [Dataset]. Data Dryad. https://doi.org/10.5061/dryad.5pt92.

ACKNOWLEDGEMENTS

We thank our collaborators at the Zooniverse and the thousands of citizen scientists who have participated in our research, particularly those who help moderate Snapshot Safari projects. We acknowledge the Minnesota Supercomputing Institute (www.msi.umn.org) for providing resources contributing to data storage and processing. We also thank our partner organizations, many of whom provided funding and promotion for this work: Cincinnati Zoo & Botanical Gardens, Detroit Zoological Society, Living Desert Zoo, Nelson Mandela University, Ruaha Carnivore Project, Seneca Park Zoo, Singita Grumeti Fund, South African National Biodiversity Institution, South African National Parks, University of Wyoming, and Zoo Miami.

FUNDING INFORMATION

This work was supported by AZA's Saving Animals from Extinction, Disney's "Protect the Pride," Microsoft's "AI for Earth," National Geographic Explorer grants, U.S. National Science Foundation grants: PRFB #1810586: IIS #1619177: OAC #1835530, National Aeronautics and Space Administration grant award #80 NSSC20M0057, and the National Research Foundation of South Africa. South African research permits were provided by South African National Parks PACC1403.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

S. Huebner led the preparation of this manuscript. M. Palmer classified the Serengeti gold standard sample and provided support on manuscript writing. C. Packer provided input on analyses and structure. All other contributions shared equally.

AUTHOR AFFILIATIONS

Sarah E. Huebner orcid.org/0000-0001-5682-6467
Smithsonian Conservation Biology Institute, US
Meredith S. Palmer orcid.org/0000-0002-1416-1732
Princeton University, US
Craig Packer orcid.org/0000-0002-3939-8162
University of Minnesota, US

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., et al. (2016) {TensorFlow}: A System for {Large-Scale} Machine Learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283.

Ackerman, S., Dube, P., Farchi, E., Raz, O., Zalmanovici, M. (2021). Machine Learning Model Drift Detection Via Weak Data Slices. In: 2021 IEEE/ACM Third International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest)., pp. 1–8. DOI: https://doi.org/10.1109/DeepTest52559.2021.00007

Ahumada, J.A. and **Fegraus, E.** (2019) Wildlife Insights: A platform to process, manage, analyze, understand and share biodiversity information from in-situ passive sensors. In: *AGU Fall Meeting Abstracts* (Vol. 2019, pp. B13 A-05 W).

Anderson, T.M., White, S., Davis, B., Erhardt, R., Palmer, M., Swanson, A., Kosmala, M., et al. (2016). The spatial distribution of African savannah herbivores: species associations and habitat occupancy in a landscape context. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1703), p. 20150314. DOI: https://doi.org/10.1098/rstb.2015.0314

Battu, T., Reddy Lakshmi, D.S. (2023). Animal image identification and classification using deep neural networks techniques. *Measurement: Sensors* 25, p. 100611. DOI: https://doi.org/10.1016/j.measen.2022.100611

Beery, S., Morris, D., Yang, S. (2019). Efficient Pipeline for Camera Trap Image Review. arXiv preprint arXiv:1907.06772. DOI: https://doi.org/10.48550/arXiv.1907.06772

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. DOI: https://doi.org/10.1007/978-3-030-01270-0_28

Bonebrake, T.C., Brown, C.J., Bell, J.D., Blanchard, J.L., Chauvenet, A., Champion, C., Chen, I.-C., et al. (2018) Managing consequences of climate-driven species redistribution requires integration of ecology, conservation and social science. *Biological Reviews*, 93(1), pp. 284–305. DOI: https://doi.org/10.1111/brv.12344

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E. (2022) Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13(8), pp. 1640–1660. DOI: https://doi.org/10.1111/2041-210X.13901

Burton, A.C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J.T., Bayne, E., et al. (2015) Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology* 52(3), pp. 675–685. DOI: https://doi.org/10.1111/1365-2664.12432

Caravaggi, A., Burton, A.C., Clark, D.A., Fisher, J.T., Grass, A., Green, S., Hobaiter, C., et al. (2020) A review of factors to consider when using camera traps to study animal behavior to inform wildlife ecology and conservation. *Conservation Science and Practice* 2(8), p. e239. DOI: https://doi.org/10.1111/csp2.239

Ceballos, G., Ehrlich, P.R. and **Dirzo, R.** (2017) Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the national academy of sciences*, 114(30), pp. E6089–E6096. DOI: https://doi.org/10.1073/pnas.170494911

Clarfeld, L.A., Sirén, A.P.K., Mulhall, B.M., Wilson, T.L., Bernier, E., Farrell, J., Lunde, G., et al. (2023) Evaluating a tandem human-machine approach to labelling of wildlife in remote camera monitoring. *Ecological Informatics*, 77, p. 102257. DOI: https://doi.org/10.1016/j.ecoinf.2023.102257

Davis, R.S., Gentle, L.K., Mgoola, W.O., Stone, E.L., Uzal, A., Yarnell, R.W. (2023) Using camera trap bycatch data to assess habitat use and the influence of human activity on African elephants (*Loxodonta africana*) in Kasungu National Park, Malawi. *Mammalian Biology*, 103, pp. 121–132. DOI: https://doi.org/10.1007/s42991-022-00330-7

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), pp. 563–576. DOI: https://doi.org/10.1093/biosci/biy068

Gadsden, G.I., Malhotra, R., Schell, J., Carey, T., Harris, N.C. (2021) Michigan ZoomIN: Validating crowd-sourcing to identify mammals from camera surveys. *Wildlife Society Bulletin*, 45(2), pp. 221–229. DOI: https://doi.org/10.1002/wsb.1175

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. DOI: https://doi.org/10.1109/CVPR.2016.90

Hofmeester, T.R., Young, S., Juthberg, S., Singh, N.J., Widemo, F., Andrén, H., Linnell, et al. (2020) Using by-catch data from wildlife surveys to quantify climatic parameters and timing of phenology for plants and animals using camera traps. *Remote Sensing in Ecology and Conservation,* 6(2), pp. 129–140. DOI: https://doi.org/10.1002/rse2.136

Hsing, P.-Y., Bradley, S., Kent, V.T., Hill, R.A., Smith, G.C., Whittingham, M.J., Cokill, J., et al. (2018) Economical crowdsourcing for camera trap image classification. *Remote Sensing in Ecology and Conservation* 4(4), pp. 361–374. DOI: https://doi.org/10.1002/rse2.84

Islam, S.B. and **Valles, D.** (2020) Identification of wild species in Texas from camera-trap images using deep neural network for conservation monitoring. In: *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*. pp. 0537–0542. IEEE DOI: https://doi.org/10.1109/CCWC47524.2020.9031190

Jones, F.M., Allen, C., Arteta, C., Arthur, J., Black, C., Emmerson, L.M., Freeman, R., et al. (2018) Time-lapse imagery and volunteer classifications from the Zooniverse Penguin Watch project. *Scientific Data*, 5(1), pp. 1–13. DOI: https://doi.org/10.1038/sdata.2018.124

Kays, R., Tilak, S., Kranstauber, B., Jansen, P.A., Carbone, C., Rowcliffe, M.J., Fountain, T., et al. (2010) Monitoring wild animal communities with arrays of motion sensitive camera traps. arXiv preprint arXiv:1009.5718. DOI: https://doi.org/10.48550/arXiv.1009.5718

Laporte-Devylder, L., Ulvund, K.R., Rød-Eriksen, L., Olsson, O., Flagstad, Ø., Landa, A., Eide, N.E., et al. (2023) A camera trap-based assessment of climate-driven phenotypic plasticity of seasonal moulting in an endangered carnivore. *Remote Sensing in Ecology and Conservation*, 9(2), pp. 210–221. DOI: https://doi.org/10.1002/rse2.304

Liu, L., Mou, C., and **Xu, F.** (2024) Improved Wildlife Recognition through Fusing Camera Trap Images and Temporal Metadata. *Diversity*, 16(3), p. 139. DOI: https://doi.org/10.3390/d16030139

Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N., Clune, J., 2021. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12(1), pp. 150–161. DOI: https://doi.org/10.1111/2041-210X.13504

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M., Packer, C., and **Clune, J.** (2017) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), E5716-E5725. DOI: https://doi.org/10.1073/pnas.1719367115

Miao, Z., Liu, Z., Gaynor, K.M., Palmer, M.S., Yu, S.X. and **Getz, W.M.** (2021) Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3(10), pp. 885–895. DOI: https://doi.org/10.1038/s42256-021-00393-0

Pacifici, M., Foden, W.B., Visconti, P., Watson, J.E.M., Butchart, S.H.M., Kovacs, K.M., Scheffers, B.R., et al. (2015) Assessing species vulnerability to climate change. *Nature Climate Change*, 5(3), pp. 215–224. DOI: https://doi.org/10.1038/nclimate2448

Palmer, M., Huebner, S., Willi, Marco, C., Fortson, L., and **Packer, C.** (2021) Citizen science, computing, and conservation: how can " Crowd AI " change the way we tackle large-scale ecological challenges? *Human Computation*, 8(2), pp. 54–75. DOI: https://doi.org/10.15346/hc.v8i2.123

Palmer, M.S., Gaynor, K.M., Becker, J.A., Abraham, J.O., Mumma, M.A., and **Pringle, R.M,** (2022) Dynamic landscapes of fear: understanding spatiotemporal risk. *Trends in Ecology & Evolution*, 37(10), pp. 911–925. DOI: https://doi.org/10.1016/j.tree.2022.06.007

Pantazis, O., Bevan, P., Pringle, H., Ferreira, G.B., Ingram, D.J., Madsen, E., Thomas, L., et al. (2024) Deep learning-based ecological analysis of camera trap images is impacted by training data quality and size. arXiv preprint arXiv:2408.14348. DOI: https://doi.org/10.48550/arXiv.2408.14348

Pardo, L.E., Bombaci, S., Huebner, S.E., Somers, M.J., Fritz, H., Downs, C., … and **Venter, J.A.** (2021) Snapshot Safari: A large-scale collaborative to monitor Africa's remarkable biodiversity. *South African Journal of Science*, 117(1–2), pp. 1–4. DOI: https://doi.org/10.17159/sajs.2021/8134

Schlaepfer, M.A. and **Lawler, J.J.** (2023) Conserving biodiversity in the face of rapid climate change requires a shift in priorities. *Wiley Interdisciplinary Reviews: Climate Change*, 14(1), p. e798. DOI: https://doi.org/10.1002/wcc.798

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

Shahinfar, S., Meek, P. and **Falzon, G.** (2020) "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *Ecological Informatics*, 57, p. 101085. DOI: https://doi.org/10.1016/j.ecoinf.2020.101085

Simpson, R., Page, K.R. and **De Roure, D.** (2014) Zooniverse: observing the world's largest citizen science platform. In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 1049–1054. DOI: https://doi.org/10.1145/2567948.2579215

Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J.T., Burton, C., … and **Rich, L.N.** (2017) Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, 15(1), pp. 26–34. DOI: https://doi.org/10.1002/fee.1448

Swanson, A., Kosmala, M., Lintott, C. and **Packer, C.** (2016) A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology*, 30(3), pp. 520–531. DOI: https://doi.org/10.1111/cobi.12695

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A. and **Packer, C.** (2015) Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific data*, 2(1), pp. 1–14. DOI: https://doi.org/10.1038/sdata.2015.26

Trouille, L., Lintott, C.J. and **Fortson, L.F.** (2019) Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proceedings of the National Academy of Sciences*, 116(6), pp. 1902–1909. DOI: https://doi.org/10.1073/pnas.1807190116

Vélez, J., McShea, W., Shamon, H., Castiblanco-Camacho, P.J., Tabak, M.A., Chalmers, C., Fergus, P., et al. (2023) An evaluation of platforms for processing camera-trap data using artificial intelligence. *Methods in Ecology and Evolution*, 14(2), pp. 459–477. DOI: https://doi.org/10.1111/2041-210X.14044

Villa, A.G., Salazar, A. and **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, pp. 24–32. DOI: https://doi.org/10.1016/j.ecoinf.2017.07.004

Westworth, S.O.A., Chalmers, C., Fergus, P., Longmore, S.N., Piel, A.K. and **Wich, S.A.** (2022) Understanding external influences on target detection and classification using camera trap images and machine learning. *Sensors*, 22(14), p. 5386. DOI: https://doi.org/10.3390/s22145386

Wevers, J., Beenaerts, N., Casaer, J., Zimmermann, F., Artois, T. and **Fattebert, J.** (2021) Modelling species distribution from camera trap by-catch using a scale-optimized occupancy approach. *Remote Sensing in Ecology and Conservation*, 7(3), pp. 534–549. DOI: https://doi.org/10.1002/rse2.207

Whytock, R.C., Świeżewski, J., Zwerts, J.A., Bara-Słupski, T., Koumba Pambo, A.F., Rogala, M., Bahaa-el-din, L., Boekee, K., et al. (2021) Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, 12(6), pp. 1080–1092. DOI: https://doi.org/10.1111/2041-210X.13576

Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., et al. (2019) Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1), pp. 80–91. DOI: https://doi.org/10.1111/2041-210X.13099

TO CITE THIS ARTICLE:

Huebner, SE, Palmer, MS and Packer, C. 2024. Human Supervision is Key to Achieving Accurate AI-assisted Wildlife Identifications in Camera Trap Images. *Citizen Science: Theory and Practice,* 9(1): 38, pp. 1–12. DOI: https://doi.org/10.5334/cstp.752

Submitted: 10 March 2024 **Accepted:** 07 October 2024 **Published:** 09 December 2024

Citizen Science: Theory and Practice is a peer-reviewed open access journal published by Ubiquity Press.