Received: 25 July 2024 | Accepted: 24 June 2025

DOI: 10.1111/2041-210X.70132

Methods in Ecology and Evolution

RESEARCH ARTICLE

Essential tools but overlooked bias: Artificial intelligence and citizen science classification affect camera trap data

Simone Santoro[1,2] | Santiago Gutiérrez-Zapata[1] | Javier Calzada[1,3] | Nuria Selva[3,4,5] | Diego Marín-Santos[6] | Sara Beery[7] | Kate Brandis[8] | Iñaki Fernández de Viana[9] | Paul Meek[10] | Alessio Mortelliti[11] | Eloy Revilla[5] | Jon Paul Rodríguez[12] | Lenka Straková[6] | Simone Tenan[13] | Manuel Emilio Gegúndez[6]

[1]Department of Integrated Sciences, Faculty of Experimental Sciences, University of Huelva, Huelva, Spain; [2]Department of Natural Sciences and Environmental Health, University of South-Eastern Norway, Bø, Norway; [3]Center for Advanced Studies in Physics, Mathematics, and Computing, University of Huelva, Huelva, Spain; [4]Institute of Nature Conservation, Polish Academy of Sciences, Kraków, Poland; [5]Doñana Biological Station, Spanish National Research Council, Seville, Spain; [6]Science and Technology Research Centre, Universidad de Huelva, Huelva, Spain; [7]Faculty of AI and Decision Making, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; [8]Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia; [9]Department of Information Technologies, Universidad de Huelva, Huelva, Spain; [10]Vertebrate Pest Research Unit, NSW Department of Primary Industries, Coffs Harbour, New South Wales, Australia; [11]Department of Life Sciences, University of Trieste, Trieste, Italy; [12]IUCN Species Survival Commission, Instituto Venezolano de Investigaciones Científicas (IVIC) and Provita, Caracas, Venezuela and [13]National Research Council, Institute of BioEconomy (CNR-IBE), San Michele all'Adige (Trento), Italy

Correspondence
Simone Santoro
Email: simone.santoro@usn.no

Funding information
Biodiversa+, Grant/Award Number: 101052342; Fundaci ón Biodiversidad; Universidad de Huelva, Grant/Award Number: FEDER UHU -202028; Agencia Estatal de Investigación, Grant/Award Number: PCI2023-145963-2; National Science Centre, Grant/Award Number: 2023/05/Y/NZ8/00104; Research Council of Norway; German Research Foundation; Fundación Biodiversidad, Ministerio para la Transición Ecológica y el Reto Demográfico

Handling Editor: Ruth Oliver

Abstract

1. Camera trapping generates vast image datasets requiring classification before downstream ecological inference, yet the influence of classification errors on subsequent analyses is often overlooked. Classification performance can vary widely depending on the classification method (e.g. citizen science vs. artificial intelligence [AI]), species, illumination conditions (diurnal vs. nocturnal) and other contextual factors.
2. We compared a citizen science classification method to two AI classifiers (EfficientNet and DeepFaune) using an expert-labelled hold-out of 51,588 images across seven classes ('empty', 'human', 'cervid', 'wild boar', 'red fox', 'leporid' and 'European badger') captured day and night. For each class and method, we quantified precision (accuracy of positive predictions) and recall (ability to detect all positive instances), then fitted single-season occupancy models to the classified data and compared estimates against expert-derived benchmarks. Finally, we conducted a large-scale simulation to investigate how true occupancy, detection probability and classification performance (recall and precision) collectively influence the accuracy (root mean square error [RMSE]) of occupancy estimates.
3. Citizen scientists exhibited consistently high precision but more variable recall. The AI classifiers outperformed the citizen science method in recall for several species, including wild boar, leporid and European badger. Both approaches performed worse on nocturnal images and showed reduced precision for night-time

'empty' images. Bias in occupancy estimates differed across species, methods and space—the AI-based estimates were generally more biased, with both the magnitude and direction of bias varying spatially, especially for rarer species such as leporids. In our simulation study, precision emerged as the strongest predictor of occupancy model accuracy, with lower precision substantially increasing RMSE. Lower occupancy rates increased RMSE, and precision regulated the impact of detection probability: at low precision, higher detection probability worsened errors; at high precision, RMSE remained low—or even decreased—as detection probability rose.

4. Although AI classifiers offer unmatched processing speed, our findings show that citizen science can reduce classification errors. Moreover, low precision and poor recall, especially for rare or nocturnal species, can substantially bias occupancy models. Based on our results, we recommend improving precision and accounting for classification quality and uncertainty to ensure robust inference from camera trap data.

KEYWORDS
artificial intelligence, camera trap, citizen science, computer vision, convolutional neuronal networks, deep learning, image classification, wildlife monitoring

1 | INTRODUCTION

Tackling the ongoing biodiversity crisis requires accurate knowledge of communities' ecological integrity, dynamics and population trends. Conservation efforts should focus on collecting current, reliable data that will inspire actions to improve ecosystem health. According to this framework, the concept of Essential Biodiversity Variables is gaining traction for rapidly diagnosing biodiversity changes (Kissling et al., 2018). However, they require cost-effective methods to characterise animal populations in large spatio-temporal studies, enabling well-informed global policies (Jetz et al., 2019).

Traditional wildlife surveying techniques, such as transect and point counts, live trapping and radio-tracking, have historically faced challenges due to logistical and economic constraints, as well as limitations in scalability (Burghardt et al., 2012). While these methods have contributed valuable insights into wildlife ecology, technological advancements in non-invasive remote-sensing methods, such as camera trapping, have revolutionised wildlife monitoring practices over the past decade (O'Connell et al., 2011). Camera traps are mostly motion-activated devices that automatically and remotely photograph targets, typically wild animals. For that aim, they employ a passive infrared sensor to detect a difference in surface temperature between animals and their environment, which triggers the camera accordingly (Welbourne et al., 2016). Technological advancement and lower sales prices have favoured a generally increased use of these devices, which have become indispensable for studying species abundance and distribution (Whytock et al., 2021) and animal behaviour (Burton et al., 2015; Caravaggi et al., 2017). Camera trapping has seen significant improvements in photo quantity, memory capacity, battery life, camera size and the inclusion of video with audio capabilities (Glover-Kapfer et al., 2019; Meek et al., 2014), positioning it as a standard method in wildlife monitoring and research and providing unparalleled opportunities for understanding wildlife populations and ecosystems. Recent advancements in camera trapping and image classification technologies have improved the reliability and cost-effectiveness of ecological research, particularly for mammals (Delisle et al., 2021). However, challenges remain in obtaining accurate data on ecological processes such as abundance, distribution and species interactions due to variations in species characteristics, camera models and settings (Hofmeester et al., 2019), as well as in the image classification process.

Despite the significant challenge in classifying camera trap images, novel citizen science platforms and machine learning techniques are revolutionising wildlife studies by facilitating image classification and deriving ecological data (McClure et al., 2020; Swanson et al., 2016). Currently, there are three primary methods for extracting species data from camera trap images: (i) classification by experts, that is, manually by the team of researchers carrying out the study, (ii) classification by volunteers in citizen science projects and (iii) automatic classification using artificial intelligence (AI). While experts may achieve the highest level of accuracy compared to citizen science and AI, the considerable amount of expert time required for image classification, as well as the associated costs, presents a notable impediment that can significantly delay the outputs and their contribution to conservation efforts (Gibbon et al., 2015; Green et al., 2020). Citizen science provides a substantially larger workforce than a standard-size team of researchers, allowing for the parallelised and efficient classification of many images. Although the

growing number of citizen science projects indicates their relevance for mammal monitoring (Swanson et al., 2016; Townsend et al., 2021), concerns persist about the time-consuming nature of training and maintaining these projects for accurate image classification. A recent study found that citizen science labelling accuracy differed among species, with more common and visually striking species (giraffe *Giraffa camelopardalis*, porcupine *Hystrix africaeaustralis*, male lion *Panthera leo* and waterbuck *Kobus ellipsiprymnus*) achieving better results than rare species (Swanson et al., 2016). However, well-designed protocols for citizen science training and data extraction have been shown to reduce classification error and, for certain species, produce labels of comparable quality to those of experts (Bird et al., 2014). Finally, the increasing use of AI in the last few years has significantly enhanced the efficiency of classifying camera trap images. However, perspectives on the use of machine learning for ecological inference vary. For instance, Whytock et al. (2021) suggested that AI classifications could often be directly used for ecological inference, bypassing manual validation. Conversely, others have highlighted that species-level misclassifications can lead to incorrect occupancy estimates (Lonsinger et al., 2024), stressing the importance of accounting for classification uncertainty (Cowans et al., 2024).

Our study aim is to evaluate the variability in classification performance among different methods: citizen science (CS-Zoo), hosted on Zooniverse (an open online platform for crowdsourced citizen science), an image-oriented Convolutional Neural Network (CNN) trained by us (EfficientNet-B5), which classifies the entire image, and an object-oriented CNN (DeepFaune, Rigoudy et al., 2023), which detects individual objects within the image and classifies them. We evaluated these methods across seven image classes, including 'empty', 'human' and mammal taxa (species and families), under various lighting conditions using a dataset of 51,588 images captured by 35 camera traps in Doñana National Park (SW Spain). Then, using single-season occupancy analyses (Mackenzie et al., 2002), we compared the occupancy predictions for the five animal classes derived from each classification method with those obtained from expert-labelled data.

2 | MATERIALS AND METHODS

All analyses were conducted in R version 4.4.1 (R Core Team, 2024) within a fully reproducible *renv* (Ushey & Wickham, 2025) environment. The Supporting Information provides complete scripts, data and outputs to reproduce every result in the manuscript, plus tutorial vignettes demonstrating how to: (1) evaluate classifier performance (recall, precision, etc.) against a known-truth subset and (2) simulate combinations of occupancy, detection, recall and precision to assess their effects on occupancy analyses. Tutorials use toy datasets for quick demonstrations, while appendix scripts include full analyses—some computationally intensive—with pre-generated outputs so results are immediately accessible without rerunning lengthy workflows.

Camera trap fieldwork was conducted under annual permits issued by the regional environmental authority of the Junta de Andalucía (permit numbers 202099900394570, 202199900279740, 2022107300000122 and 2023107300001409). This study did not involve human participants or personal data; therefore, approval from a human-research ethics committee was not required. Ethical approval was not required for this study, as it involved only non-invasive camera trapping and no handling or capture of animals.

2.1 | Study area and image collection

We obtained camera trap images from an ongoing wildlife monitoring project in the Doñana National Park, a UNESCO World Heritage Site in southwestern Spain. The Park's diverse ecosystems, ranging from marshlands and Mediterranean forests to coastal dunes, support a diverse mammal community. Currently, the medium and large mammal community in Doñana comprises three species of wild ungulates (red deer, *Cervus elaphus*; fallow deer, *Dama dama*; and 'wild boar', *Sus scrofa*), five species of carnivores (Iberian lynx, *Lynx pardinus*; 'red fox', *Vulpes vulpes*; European badger, *Meles meles*; Egyptian mongoose, *Herpestes ichneumon*; and genet, *Genetta genetta*), and two species of lagomorphs (Iberian hare, *Lepus granatensis*; and European rabbit, *Oryctolagus cuniculus*). Additionally, domestic ungulates such as cattle and horses are present in the area. The Iberian lynx, an endemic and endangered feline species, specialises in preying on European rabbits, a keystone species that has experienced significant population declines due to disease outbreaks. Thirty-five camera traps were randomly distributed in the study area (Figure 1), starting from October 2020. These traps were installed on wooden poles positioned 50 cm above the ground and spaced at least 1 km apart.

We set up the cameras to take three consecutive photos, with a 1-s delay between bursts. We checked the camera statuses and replaced the memory cards periodically, approximately every month. We used no-glow (black or invisible) flash cameras, specifically the Browning Dark OPS HD PRO X and Browning Dark OPS PRO DCL models. These cameras emit wavelengths below 940 nm, making them invisible to humans and mammals. However, their nocturnal images are black-and-white and may have lower quality than those captured with white flashes.

2.2 | Image classification through citizen science

We obtained image classifications through a Zooniverse CS project (CS-Zoo) that we launched. Zooniverse is the world's most popular online platform for citizen science, with over one million volunteers supporting research projects across various topics. This platform allows researchers to create a personalised website where volunteers receive information and tools to help classify subjects of different natures. In recent years, the number of camera trap projects hosted at Zooniverse has increased significantly, outpacing the

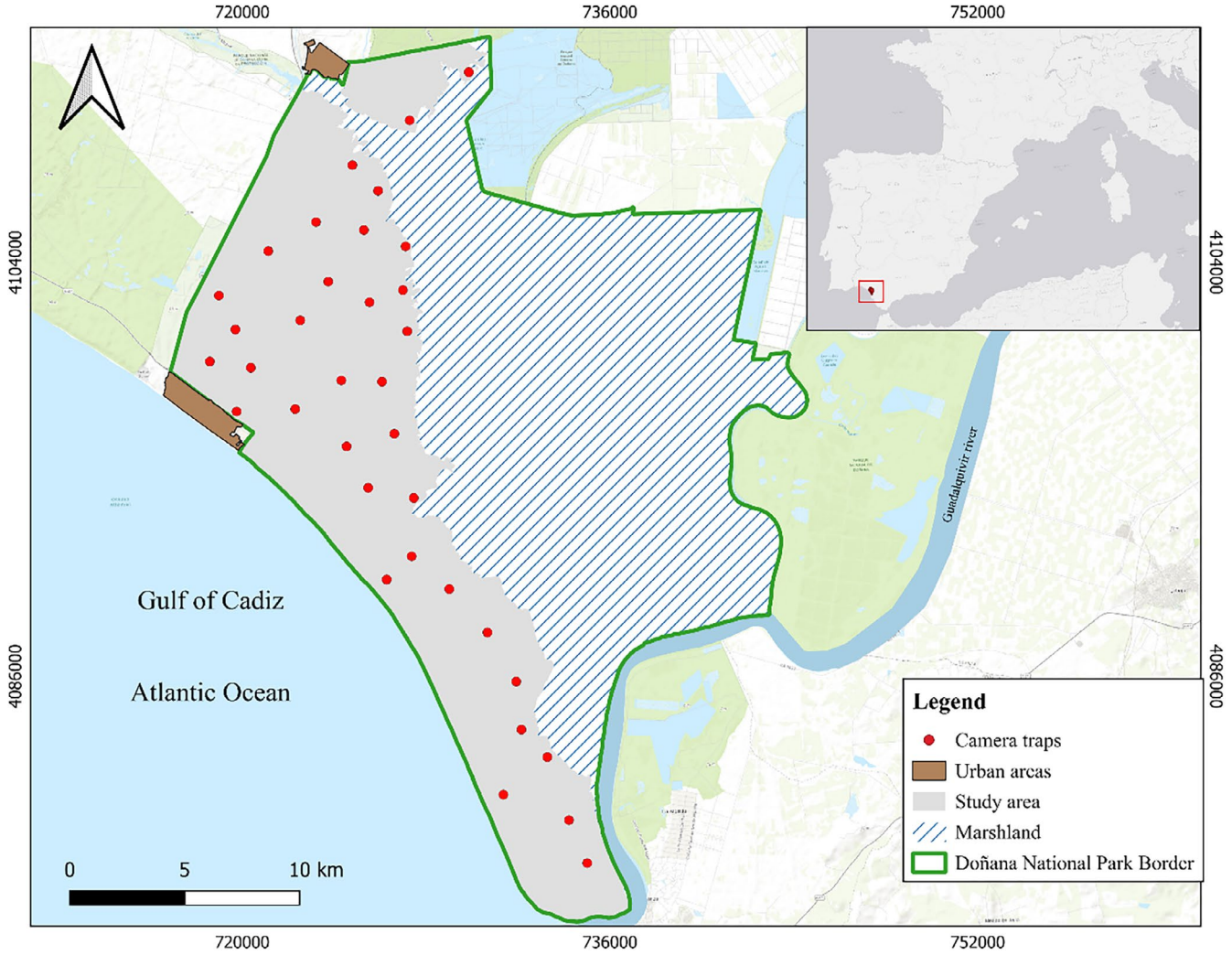

FIGURE 1 Map of the study area, the Doñana National Park in southwestern Spain, highlighting camera trap (red points) placements within the non-flooded zone (solid area). The dashed area represents the flooded zone.

growth of registered volunteers (Swanson et al., 2016; Townsend et al., 2021). Our citizen science project has enlisted about 18,000 volunteers, who, from April 2021 to January 2024, have collectively classified approximately 1,100,000 images. Volunteers were given extensive didactic material, which included photographs accompanied by explanatory text, to help them distinguish various species and identify them within the photos.

After each monthly field inspection of the camera traps, we sent a random selection of images, with a maximum of five per event (defined as consecutive photos from the same camera trap location within 90 s), to the Zooniverse project, ensuring coverage of all camera traps. Volunteers classified each image individually without viewing the sequences. Their task was to label each image with the correct species class among possible options (see below and Figure 2a). The interface provided a reference image and a warning about potential classes that could be easily confused to aid their decision-making (Figure 2b). Finally, volunteers were also requested to indicate the number of individuals observed in each image.

Volunteers were tasked with selecting from a range of 18 distinct classes to classify each image. This set encompassed 12 mammal species: the Iberian lynx, common genet, mongoose, 'red fox', 'European badger', red deer, fallow deer, 'wild boar', European rabbit, Iberian hare, cow and horse. Additionally, two mammal families, *Cervidae* (including red and fallow deer, 'cervid' class hereon) and *Leporidae* (including rabbits and hares, 'leporid' class hereon), were included in the set. These last two classes were added because, in some circumstances (low light, animals in motion or far away from the camera), it is very difficult or almost impossible to distinguish between the red deer and the fallow deer or the Iberian hare and the European rabbit. Still, it is possible to determine the family, that is, whether the photographed animal is a 'cervid' or a 'leporid'. In other circumstances where it was not feasible to identify the animal captured in the image, volunteers were allowed to classify the image as 'unrecognisable'. Finally, three additional classes, namely 'empty', 'other species' and 'human', were also included in the classification options. Labellers were instructed to classify instances

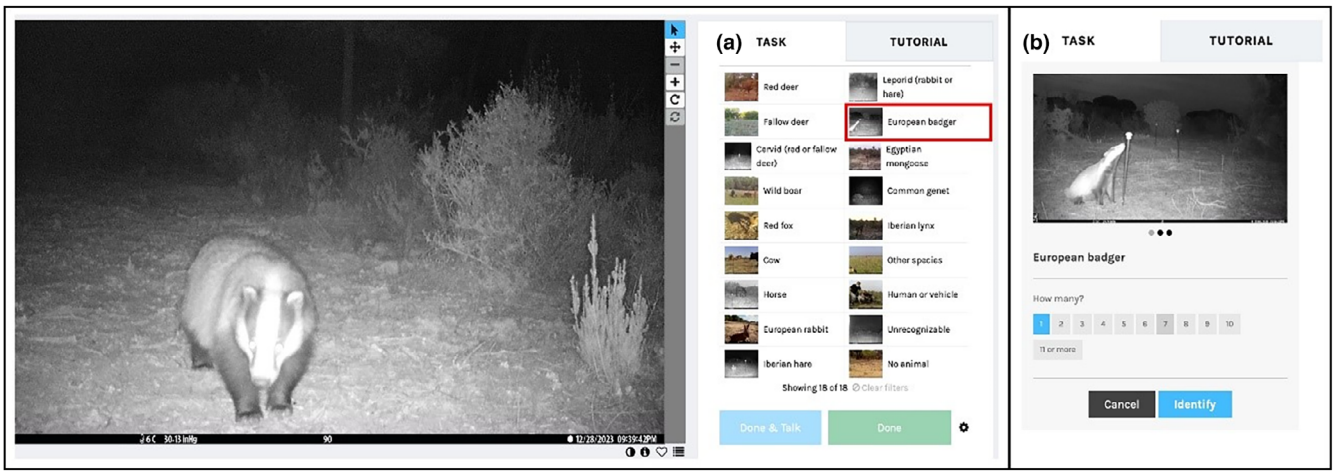

FIGURE 2 Citizen science project interface for image classification. (a) Volunteers were first asked to recognise one among the 18 possible classes: 12 mammal species, two mammal families (Cervidae and Leporidae), human or vehicle, unrecognisable, other species and no animal (i.e. 'empty'). (b) Next, if a mammal class was selected, they were asked to indicate how many individuals they saw.

of humans or vehicles as a single class. Consequently, throughout the article, the class 'human' encompassed humans and vehicles.

We established specific criteria for removing images from the set available for volunteer classification at Zooniverse. An image was retired from volunteers' classification and considered as classified if it met any of the following conditions: (i) two or more volunteers identified it as 'human', (ii) three or more volunteers classified it as 'empty' or (iii) five or more volunteers reached a consensus on the same classification. Additionally, once an image received 15 classifications, it was retired from the platform. If a majority class emerged among these 15 classifications, the image was stored with that classification. If no majority emerged, the image was assigned as NA (not classified by citizen science).

2.3 | Image classification through AI

A CNN classification model was developed with the *EfficientNet-B5* (EfficientNet hereon), an architecture designed to optimise model performance and computational resources (Tan & Le, 2019). The network was built and trained to classify 12 mammal species, four taxonomic groups (cervid, leporid, small mammal, bird), 'empty' and 'human' classes. However, unlike Zooniverse, the network lacked the 'unrecognisable' class. During the training phase, a dataset comprising 390,208 training images and 7200 validation images (about 70% from our study area but different from the set used for performance evaluation) was utilised, with the latter evenly distributed across all classes, containing 400 samples per category. We employed transfer learning techniques by initialising the model's weights with pre-trained weights from the ImageNet dataset. This approach facilitated the model's learning process and convergence during training (Russakovsky et al., 2015; Torrey & Shavlik, 2010). Furthermore, we applied data augmentation techniques to expand the dataset and enhance the model's robustness against input data variations. With images standardised to a resolution of 480×640 pixels, augmentation strategies such as rotation, flipping and scaling were systematically employed to diversify the dataset and imbue the model with greater resilience to real-world scenarios (Shorten & Khoshgoftaar, 2019). To mitigate class imbalances and promote equitable representation across all categories, at each epoch of the training phase, 1160 images per class were randomly sampled with replacement, aligning with the size of the minority class (dog class). This approach ensured that each class contributed equivalently to the model's learning process, thereby minimising the risk of bias and maximising classification accuracy, particularly for underrepresented categories.

A second AI classification was obtained using DeepFaune v1.2 (https://www.deepfaune.cnrs.fr). DeepFaune employs YOLOv8s, a high-performance, medium-sized detection model trained using cropping data from MegaDetector V6 (Beery et al., 2018). DeepFaune's alternative detector utilises the regions defined by the detection of each animal to provide classification within those regions rather than classifying the entire image. It identifies empty images, as well as people, vehicles and 28 mammal species or higher taxonomic groups trained in the CNN classifier. In contrast, EfficientNet does not perform image cropping. We specifically tailored the selection of mammal species to be recognised by DeepFaune to match those existing in our study area. DeepFaune users are required to set a confidence threshold. We set this threshold to the minimum allowed value (0.25), which closely matched the minimum confidence score of our EfficientNet model (0.23).

2.4 | Classification performance dataset

The original dataset comprised 55,059 images captured between September 2021 and January 2022. These images constituted a comprehensive sample from our citizen science project, representing a random selection from the entire pool of images collected during the study period. None of these images was employed to train the AI classification systems. Ground-truth labels for these images were

reviewed by three mammalogists involved in this study. In cases of uncertainty, experts reviewed the entire photo sequence, benefiting from a level of scrutiny not available to the CS-Zoo and AI classification systems. Additionally, instances initially labelled as 'empty' by experts but later identified differently by MegaDetector V.5 were re-evaluated by experts ($n = 1357$ images). This re-evaluation led to changes in classification in 370 cases in the expert dataset out of the 55,059 images. Seven classes were selected based on expert classification, each comprising a minimum of 100 instances: 'empty', 'human' and five species or families—'cervid', 'wild boar', 'red fox', 'leporid' and 'European badger'. The final expert-labelled dataset was unevenly distributed: 'empty' = 30,948, 'cervid' = 13,073, 'wild boar' = 1171, 'red fox' = 998, 'leporid' = 439 and 'European badger' = 129. This study dataset differed from the original dataset as it excluded images belonging to any other species from the original 18 distinct classes, focusing solely on the seven specified classes for the case study.

2.5 | Evaluation of the classification systems performance

In machine learning, four terms are commonly used to assess image classifications: true positive, true negative, false negative and false positive. True positives correctly identify the presence of the class, true negatives correctly identify the absence of the class, false positives incorrectly identify the presence of the class when it is not actually present and false negatives incorrectly fail to identify the presence of the class. Additional concepts, like recall and precision, are fundamental in this context. *Recall* (also known as *sensitivity*) measures the proportion of actual positive instances correctly identified by the classifier. It answers the question: 'Of all the photos of a given class, what proportion were correctly identified by the classifier?'. *Precision* measures the proportion of positive instances identified by the classifier that were actually correct. It answers the question: 'Of all the photos identified as a given class, what proportion actually belonged to that class?'. Maximising both recall and precision for all the classes is crucial for monitoring studies and evidence-based conservation. For example, high recall ensures most instances of a given threatened species are detected, while precision confirms accurate species identification.

We used expert classifications as the ground truth to compare the performance of three classifiers—CS-Zoo and two AI models (EfficientNet and DeepFaune)—in identifying seven classes: 'empty', 'human', 'cervid', 'wild boar', 'red fox', 'leporid' and 'European badger'. We employed three performance metrics: *recall*, *precision* and *Matthew's correlation coefficient* (Chicco & Jurman, 2022). *Matthew's correlation coefficient* provides a single score that evaluates how well the predicted classifications match the actual classes in the photos, even when there is an imbalance among the classes.

First, we evaluated the classifiers' performance for each class, irrespective of light conditions (diurnal or nocturnal). Next, we investigated the impact of light conditions on classification accuracy. We categorised each photo as 'day' or 'night' based on its timestamp relative to the official sunrise and sunset times for the study area, cross-referencing data from the National Institute of Geography (https://www.ign.es). Photos captured under ambiguous lighting conditions (dawn and dusk) were excluded from the subsequent analysis. We employed resampling techniques ($n = 1000$) with replacement (Efron & Tibshirani, 1991) to calculate 95% credible intervals for each metric and class.

2.6 | Statistical analyses

2.6.1 | Differences in classification performance

We evaluated performance differences between classifiers, classes and light conditions. To achieve this, we utilised the *glmmTMB* function (Brooks et al., 2017) from the homonymous package to run generalized linear models, regressing recall and precision onto *classifier* (CS-Zoo, EfficientNet, DeepFaune), *class* ('empty', 'human', 'cervid', 'wild boar', 'red fox', 'leporid' and 'European badger') and *light conditions* (day or night). We used a beta distribution for the response variables and checked model fit using the *simulateResiduals* function from the *DHARMa* package (Hartig, 2021). To assess interactions and isolate individual contributions, we ran three GLMs for recall and precision, testing pairwise interactions (e.g. *classifier* × *light*) while treating the third predictor (e.g. *class*) additively. Post-hoc pairwise comparisons were conducted using the *emmeans* function (Lenth, 2022), with Tukey's *p*-value adjustment for multiple comparisons. As the model follows a beta-binomial distribution, comparisons were made on the log odds ratio scale.

2.6.2 | Single-season occupancy

For the data obtained from each classifier and animal class during the initial 30 days of the study period, we applied single-season occupancy models (Mackenzie et al., 2002) using the *occu* function from the *unmarked* package (Fiske & Chandler, 2011; Kellner et al., 2023). Single-season occupancy models include two parameters: (1) occupancy (ψ), which represents the probability that the target species used a camera trap station during the season and (2) detection probability (p), which denotes the probability of detecting the target species during a survey if the species was present at the station. To ensure uniform model selection criteria and avoid biases stemming from class-specific candidate model sets, we evaluated a consistent set of 44 candidate models across 16 combinations of classifiers (ground truth, CS-Zoo, EfficientNet and Deepfaune) and animal taxon ('cervid', 'wild boar', 'red fox', 'leporid' and 'European badger'). The most parameterised global model included the non-additive effects of latitude and longitude for both occupancy and detection parameters, and the effect of survey periods (three groups of 10 consecutive days) on detection. We excluded candidate models if they exhibited convergence issues

during likelihood maximisation, non-estimable standard errors or a Hessian condition number exceeding 10^4. Next, we assessed the models' goodness of fit (using the *parboot* function). We used the smallest estimate of c-hat as the reference for each set of candidate models (Burnham & Anderson, 2002). We found no overdispersion issues, as the c-hat estimates of each set of candidate models were equal to 1 after rounding to the second decimal place. We then used the *modSel* function to create AIC tables (the default criterion in *unmarked*). We predicted the occupancy estimates in the study area by model-averaging over these models (using the *fitList* and *predict* functions).

2.6.3 | Simulation study: Effect of classification errors on occupancy estimates

We simulated occupancy data on a 100×100 grid (10,000 cells) under nine combinations of true occupancy ($\psi = 0.2, 0.5, 0.8$) and cumulative detection probability ($p = 0.2, 0.5, 0.8$)—where 'cumulative *p*' is the chance of detecting a species at least once over the five occasions of a survey. Two standardised spatial covariates (one each for occupancy—*covPsi*—and per-occasion detection—*covP*) were generated by summing 10 random Gaussian peaks to create a heterogeneous landscape of ψ and *p* values.

For each (ψ, *p*) pair, 100 bootstrap iterations were performed. In each iteration, 200 sites were sampled roughly evenly across the grid; true occupancy was drawn as $logit^{-1}[logit(\psi) + covPsi]$, and conditional on presence, a five-occasion detection history was generated with per-occasion probability $logit^{-1}[logit(p) + covP]$. A single-season occupancy model was then fitted using *unmarked* (*occu(~covP ~ covPsi)*; Fiske & Chandler, 2011; Kellner et al., 2023), and ψ was predicted for all grid cells, averaging across iterations to obtain a baseline surface.

Next, we iterated over recall and precision values ranging from 0.50 to 1.00 in increments of 0.01, where, for example, a recall of 0.50 means that only half of the true detections are retained (i.e. 50% of false negatives) and a precision of 0.5 that only half of the detections are true (i.e. 50% of false positives). For each recall and precision pair, we injected classification errors into the original detection history by randomly flipping non-detections to detections to generate false positives and flipping detections to non-detections to generate false negatives in proportions that achieve the target recall and precision. We then refitted the same occupancy model to these perturbed data. We averaged the predicted ψ over 100 iterations to obtain estimates of ψ when recall and/or precision were below one. When recall equals 1 and precision equals 1 (no misclassification), we reused the baseline ψ. From each scenario (ψ, p, recall, precision), we calculated the RMSE of predicted ψ against the true ψ surface. After standardising the true ψ, p, recall and precision (mean = 0, SD = 1), we fitted a Gamma-family GLM (log link) predicting RMSE as a function of all four standardised covariates and their two-way interactions. Model fit was assessed via *DHARMa* (Hartig, 2021), and effect estimates (95% CI) were extracted with *effects* (Fox & Weisberg, 2019).

3 | RESULTS

When the study dataset was created, our citizen science project in the Zooniverse platform had been active for approximately 443 days, classifying about 677,000 images, indicating an average classification rate of 1528 daily. The EfficientNet model ran on a computer with an Intel(R) Core(TM) i7-5820K CPU @ 3.30GHz and an NVIDIA GeForce GTX 1080 graphics card, processed 9 photos per second or 777,600 per day (>500 times faster than CS-Zoo). To process the same number of images as Zooniverse in 443 days, our AI model would require approximately 21 h. The confusion matrices of the three classifiers and a detailed output of statistical analyses can be found in Appendix S1.

Performance metrics varied notably across classes and classification systems (Figure 3). While the AI models consistently achieved higher recall values for four of the five animal taxa during night-time, CS-Zoo demonstrated superior precision across all combinations of classes and light conditions.

3.1 | Differences in recall

For the pooled dataset, EfficientNet detected fewer instances of the 'empty' class than both CS-Zoo (all values reported on the log odds ratio scale: -1.340 ± 0.485, $p = 0.016$) and DeepFaune (-1.372 ± 0.489, $p = 0.014$). For the 'European badger', both EfficientNet (2.597 ± 0.515, $p < 0.001$) and DeepFaune (2.353 ± 0.472, $p < 0.001$) outperformed CS-Zoo. EfficientNet also had higher recall for the 'leporid' class compared to CS-Zoo (1.199 ± 0.270, $p < 0.001$) and DeepFaune (0.867 ± 0.274, $p = 0.004$). However, DeepFaune showed significantly higher recall for 'red fox' than both EfficientNet (0.626 ± 0.242, $p = 0.026$) and CS-Zoo (0.608 ± 0.241, $p = 0.031$).

Recall differed between day and night only for CS-Zoo, which showed a significant decrease at night (-1.201 ± 0.297, $p < 0.001$). DeepFaune showed a marginal decrease (-0.548 ± 0.313, $p = 0.080$), while EfficientNet showed no significant difference between day and night. By class, recall for the 'empty' class significantly decreased at night (-2.095 ± 0.613, $p < 0.001$), and 'red fox' also showed a smaller but significant decrease (-0.803 ± 0.311, $p = 0.010$). Other species showed no significant variation, and recall estimates for 'European badger', 'human' and 'leporid' could not be estimated due to insufficient data.

3.2 | Differences in precision

For the pooled dataset, we found no differences between EfficientNet and DeepFaune. In most cases, CS-Zoo was more precise than DeepFaune (all values reported on the log odds ratio scale: 'European badger': 3.359 ± 1.030, $p = 0.003$; 'human': 3.388 ± 1.030, $p = 0.003$; 'leporid': 3.292 ± 1.030, $p = 0.004$; 'red fox': 1.872 ± 0.749, $p = 0.033$) and EfficientNet ('European badger': 2.966 ± 1.050, $p = 0.013$; 'human': 2.514 ± 1.080, $p = 0.051$; 'leporid':

2.720 ± 1.060, $p = 0.028$; 'red fox': 1.997 ± 0.746, $p = 0.020$; 'wild boar': 2.321 ± 0.783, $p = 0.009$).

The overall precision of EfficientNet and DeepFaune declined from day to night (EfficientNet: −0.998 ± 0.493, $p = 0.043$; DeepFaune: −0.976 ± 0.435, $p = 0.025$), primarily driven by a sharp decrease in the 'empty' class (−2.307 ± 0.429, $p < 0.0001$). In contrast, CS-Zoo showed no significant differences.

3.3 | Occupancy analyses

For 'cervid', all methods produced occupancy estimates similar to expert labelling, with occupancy being high and uniform across the study area (Figure 4). CS-Zoo also performed well for 'leporid' and 'European badger', with only minor overestimation in high-occupancy areas. EfficientNet and DeepFaune overestimated 'wild boar' occupancy in low-occupancy areas and underestimated it in high-occupancy areas, a pattern also observed for 'red fox'. For 'European badger', EfficientNet showed the strongest bias, underestimating occupancy in areas with higher occupancy values and overestimating it in areas with lower occupancy, a trend also observed in DeepFaune but to a lesser extent. For 'leporid', which had low true occupancy throughout the area, EfficientNet and especially DeepFaune overestimated occupancy, while CS-Zoo showed a milder bias similar to that seen for 'red fox' (Table 1).

3.4 | Effect of classification errors on occupancy model accuracy

We found that classification precision was by far the strongest driver of occupancy model accuracy: a 1 SD increase in precision reduced RMSE by 0.621 SD ($p < 0.001$; Figure 5a). True occupancy also had a significant but more moderate effect (−0.129 SD per SD increase in ψ; $p < 0.001$; Figure 5b), with higher ψ leading to lower error. Detection probability showed a slight negative main effect (−0.013 SD; $p < 0.001$) and recall a modest reduction in RMSE (−0.027 SD; $p < 0.001$).

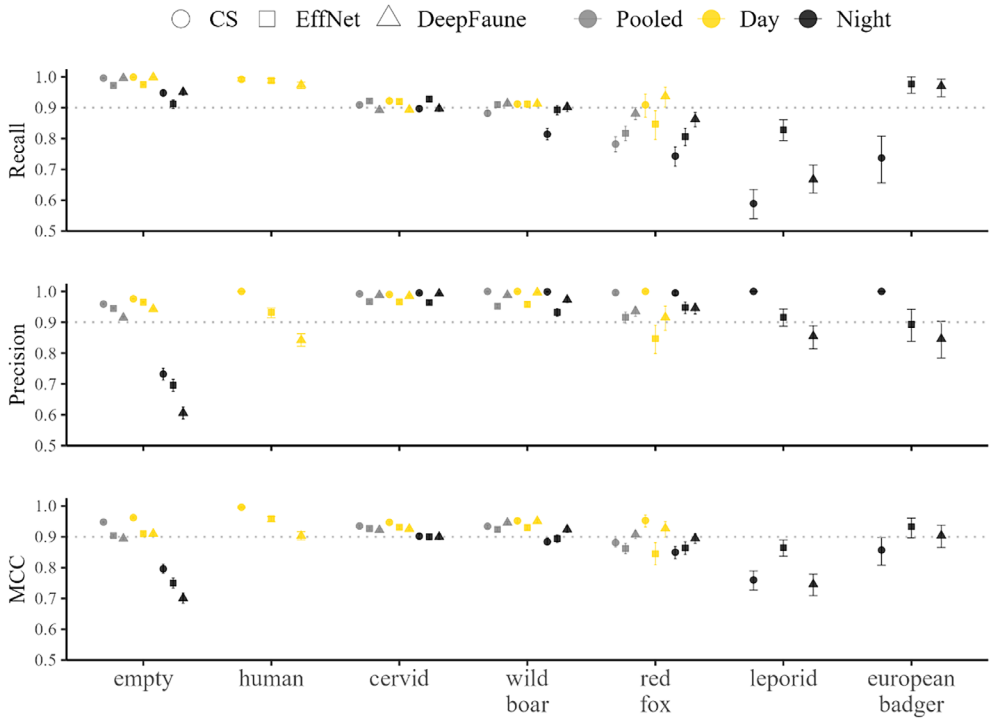

FIGURE 3 Variation in *recall*, *precision* and *Matthew's correlation coefficient* (MCC) metrics across different classes and classification systems (CS-Zoo, EfficientNet, DeepFaune) showed both pooled and separated between day and night. No metrics for nocturnal human and diurnal 'leporid' and 'European badger' are reported due to limited detection instances (<1%). Recall varied considerably among classes and classifiers. However, the AI models consistently achieved higher values for four out of the five animal taxa during night-time in recall, while CS-Zoo exhibited superior precision across all combinations of classes and light conditions. Nocturnal photos demonstrate lower overall performance (MCC), primarily due to decreased precision, particularly notable for the 'empty' class.

FIGURE 4 Comparative analysis of model-averaged predictions for occupancy probabilities estimated by CS-Zoo, EfficientNet and DeepFaune compared to expert-labelled data for 'cervid', 'wild boar', 'red fox', 'European badger' and 'leporid'. The first column represents the occupancy model-averaged predictions by expert labelling, while the subsequent columns show deviations of predictions by CS-Zoo, EfficientNet and DeepFaune. The scales represent occupancy probability (p: 0.00 to 1.00) and deviations (−1 to +1).

BRITISH ECOLOGICAL SOCIETY Methods in Ecology and Evolution

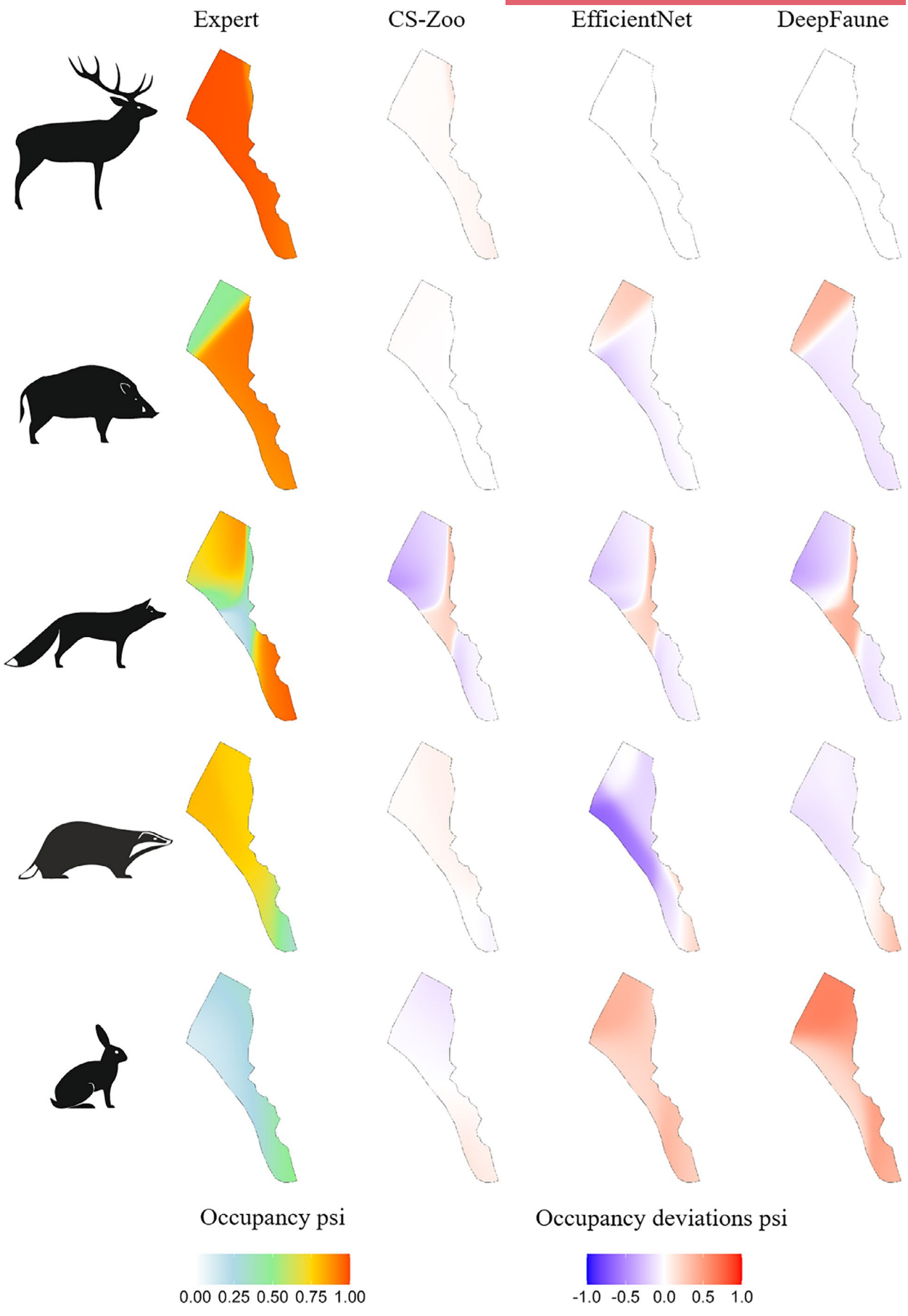

TABLE 1 Comparative analysis of model selection and model-averaged predictions for animal taxa and classification systems. Rows in bold represent the reference method (expert classification). The 'Occupancy formula' and 'Detection formula' columns represent the corresponding parameters' formulas selected from the lowest AIC model in the respective sets of candidate models for each classifier and class. The 95% credible interval (CI) estimates correspond to the median and 0.025 and 0.975 percentiles of the model-averaged predictions over the study area.

Species	Classifier	Occupancy formula	Detection formula	95% CI estimate
cervid	**Expert**	**constant**	**constant**	**0.98 (0.916, 0.984)**
	CS-Zoo	constant	constant	0.999 (0.999, 1)
	EfficientNet	constant	constant	0.98 (0.916, 0.984)
	DeepFaune	constant	constant	0.98 (0.916, 0.984)
wild boar	**Expert**	**lon + lat**	**lon + lat + lon:lat**	**0.9 (0.465, 0.936)**
	CS-Zoo	lon + lat	lon + lat + lon:lat	0.901 (0.482, 0.938)
	EfficientNet	lon + lat	lon + lat + lon:lat	0.831 (0.545, 0.918)
	DeepFaune	constant	lon + lat	0.828 (0.755, 0.881)
red fox	**Expert**	**lon + lat + lon:lat**	**lon**	**0.9 (0.465, 0.936)**
	CS-Zoo	lon	lat	0.901 (0.482, 0.938)
	EfficientNet	lon + lat + lon:lat	lon + lat + lon:lat	0.831 (0.545, 0.918)
	DeepFaune	lon	lon	0.828 (0.755, 0.881)
leporid	**Expert**	**lon**	**constant**	**0.237 (0.132, 0.475)**
	CS-Zoo	lon	lon	0.173 (0.112, 0.562)
	EfficientNet	lon + lat + lon:lat	constant	0.548 (0.351, 0.75)
	DeepFaune	lon + lat + lon:lat	constant	0.761 (0.346, 0.901)
badger	**Expert**	**lon**	lon + lat + lon:lat	**0.762 (0.4, 0.806)**
	CS-Zoo	lon	lon + lat + lon:lat	0.825 (0.363, 0.829)
	EfficientNet	lon + lat	lon + lat + lon:lat	0.54 (0.167, 0.771)
	DeepFaune	constant	lon + lat + lon:lat	0.686 (0.657, 0.729)

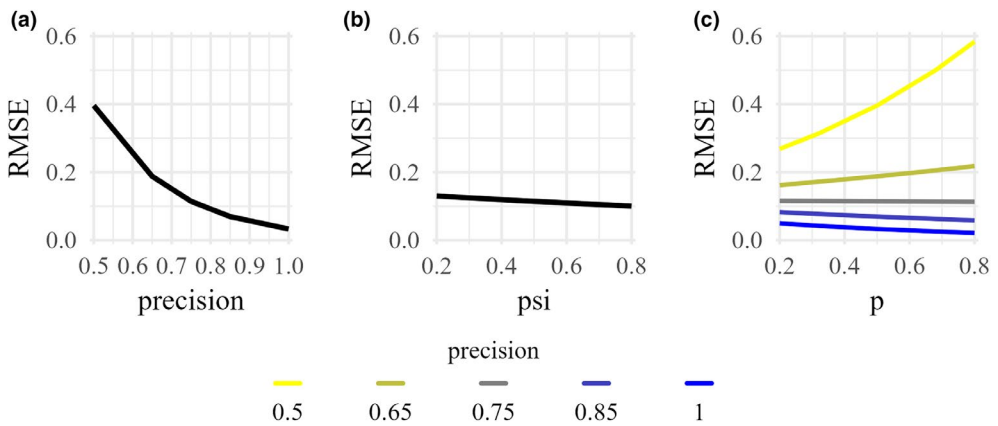

FIGURE 5 Simulation-derived partial effects of precision, true occupancy and detection probability on occupancy model accuracy (root mean square error [RMSE]), with predictions plotted on the response scale. These plots highlight the three strongest influences on RMSE: (a) higher precision drives down RMSE most steeply, (b) higher true occupancy also reduces RMSE, but to a lesser extent and (c) precision moderates the effect of detection probability—at low precision, RMSE rises with p, while at high precision RMSE remains low or declines.

Among interactions, only ψ × precision was non-significant ($p = 0.547$). In contrast, the precision × p interaction revealed that at low precision, increasing detection probability raises RMSE, whereas at high precision, RMSE stays low or even declines slightly as p increases ($p < 0.001$; Figure 5c). Other two-way interactions were statistically significant, but their effect sizes were more negligible (see Supporting Information for full model output).

4 | DISCUSSION

Evaluating classification performance in camera trap studies is challenging due to variability among classifiers, species, lighting conditions and other factors. Our citizen science project excelled in precision; however, all classifiers experienced reduced performance in nocturnal images, particularly for the 'empty' class, raising

concerns about the potential for missed detections of nocturnal species. Estimates from single-season occupancy models, which do not account for false positives, were accurate for uniformly distributed species using both AI and citizen science. However, they were significantly biased for rarer or more heterogeneous species, particularly with AI, due to lower precision. Simulations showed that, when classification precision is low, higher detection probability paradoxically increases occupancy model error by spreading false positives across a larger share of the study area. Collectively, these findings underscore the need to enhance classification precision and explicitly incorporate classification uncertainty into ecological models to ensure the reliability of automated camera trap monitoring.

Our findings revealed high classification precision in our citizen science project, consistent with the common practice of benchmarking AI model accuracy against citizen science projects, often considered the gold standard (Norouzzadeh et al., 2018; Sullivan et al., 2018; Swanson et al., 2016; Willi et al., 2019). However, we observed notable disparities in the performance of classification systems across various target classes. While all systems accurately identified common classes like 'human', 'cervid' and 'wild boar', they struggled with less common ones such as 'red fox', 'leporid' and 'European badger', a similar outcome to those reported in previous research. For instance, Swanson et al. (2016) achieved accurate classifications only for common species in a large-scale citizen science camera trap study in the Serengeti National Park. Vélez et al. (2022) noted that AI platforms like Wildlife Insights excel in accuracy for common species but have lower recall for less common ones, highlighting the difficulty of applying a single AI model across diverse ecosystems and taxa. Beery et al. (2018), Norouzzadeh et al. (2021) and Norman et al. (2023) demonstrated that AI using object detection with cropped regions around animals can reduce overfitting to specific camera locations, improving generalisability. Our study compared DeepFaune, employing object detection with cropped regions, to EfficientNet, which was trained on whole-image labels largely from our study area. While no significant performance differences were found, the two models differ in their transferability. DeepFaune, trained on cropped regions and a broader dataset, demonstrates better adaptability across diverse ecosystems. In contrast, EfficientNet, trained on whole images specific to the study area, is more influenced by habitat and background features, making it less transferable to new environments without retraining on local data.

Photos' light condition was another critical aspect influencing classification performance. Our study highlights a decline in the ability of classification systems to accurately identify the 'empty' class in night images, indicating potential underdetection of nocturnally active species. Accordingly, we observed a significant decline in classification performance across all systems for night-time photos, which was particularly evident for nocturnal species like 'leporid', 'European badger' and fast-moving species like 'red fox'. This decline deserves attention as nocturnality is widespread among mammals and many species are becoming more nocturnal to avoid human disturbance (Gaynor et al., 2018). Besides hindering movement detection (Hofmeester et al., 2019), lower identification accuracy in nocturnal images is likely due to lower image quality from infrared flashes. Notably, although no-glow flashes are invisible to humans, some animals can detect them and alter their behaviour (Meek et al., 2016), potentially affecting detection rates. Rowcliffe et al. (2014) noted a 21% increase in camera trap detection radius during the day, affecting motion detection and, subsequently, diel activity and estimation of demographic parameters based on it. Cusack et al. (2015) observed that random encounter models applied to daytime observations of lions, where movement is less random, may yield biased density estimates. These movement detection and species identification challenges can distort our understanding of species' activity patterns and generate indirect implications for estimating ecological parameters.

Misclassification at the species level introduces false negatives for the misclassified species and false positives for the species that are wrongly attributed, potentially biasing occupancy estimates if not addressed. In our single-season analyses, all methods matched reference occupancy for the common 'cervid'; however, AI overestimated occupancy in low-occupancy cells and underestimated it in high-occupancy cells for 'wild boar' and 'red fox'. 'European badger' showed a similar pattern, and both AI models greatly overestimated occupancy for the consistently rare 'leporid', whereas CS-Zoo displayed only mild bias. Simulations clarify these patterns: when precision is low, raising detection probability increases the absolute number of false positives, seeding unoccupied sites and inflating occupancy estimates, while false negatives are partly buffered within sampling intervals spanning multiple days because later true detections can offset earlier misses; a single false-positive frame, however, irreversibly tags the entire occasion as a detection, thereby helping make precision—not recall—the dominant driver of bias. Although our simulations report a mean error across the landscape, the magnitude and even direction of bias vary sharply among cells (Figure 4; Supporting Information heat maps), mirroring the spatial heterogeneity in the empirical data. Unmodelled misclassification can thus distort overall occupancy estimates and bias spatial covariate effects, a concern, especially for rare or endangered species. False-positive occupancy models offer a potential remedy (Chambert et al., 2015; MacKenzie et al., 2017; Royle & Link, 2006), though they often face estimability challenges (Cowans et al., 2024; Dussert et al., 2024; Monchy et al., 2025) and multi-species occupancy models widely used in camera trap studies (Devarajan et al., 2020) likewise ignore false positives. Enhancing classification precision and formally propagating classification uncertainty through occupancy models are, therefore, critical steps for reliable automated camera trap monitoring and robust ecological inference. Importantly, although these biases emerged in our modest study area, the same false-positive rate applied to regional or national monitoring networks would generate far more spurious detections and could thus mislead broad-scale occupancy estimates and conservation decisions.

Citizen science projects engage a diverse audience interested in quantifying biodiversity. While this study does not explore volunteers' motivations or perceptions, maintaining public involvement is

crucial. Leveraging CS data for AI algorithm training enhances classification outcomes by combining the strengths of both approaches (Fortson et al., 2012; McClure et al., 2020; Tuia et al., 2022; Willi et al., 2019). Integrating AI models and CS platforms improves species identification accuracy and reduces human labelling efforts. For instance, Norouzzadeh et al. (2018) trained a CNN model on 3.2 million images from the Snapshot Serengeti dataset, achieving accuracy comparable to trained citizen scientists. Similarly, Willi et al. (2019) integrated deep learning with CS through a Zooniverse project, enhancing species identification accuracy and expediting classification. This integration holds promise for large-scale monitoring or research, enabling rapid, precise and cost-effective data processing across broader spatial and temporal scales in ecological monitoring.

Our findings highlight the challenges of wildlife monitoring using camera traps and identify issues of particular concern. The imperfect functioning of passive infrared motion detectors introduces significant variability, and even high-quality cameras can miss many events at rates that depend on climate and species (Urbanek et al., 2019). Hofmeester et al. (2019) identified 40 factors affecting animal detection and identification via camera trapping, underscoring the method's complexity. We emphasise that the classification process is another potential source of bias in ecological inference from camera trap data. Carefully considering these factors is critical to improving data accuracy and reliability. Growing efforts are now exploring ways to explicitly incorporate classification uncertainty into ecological models—particularly occupancy models—by integrating species-level classification probabilities directly into the estimation process (Cowans et al., 2024; Dussert et al., 2024; Lonsinger et al., 2024; Monchy et al., 2025; Rhinehart et al., 2022). However, robust statistical frameworks capable of handling such uncertainty across large-scale, multi-species datasets remain under development, highlighting the need for further methodological research. Adopting multi-event (Pradel, 2005) and hierarchical modelling (Kery & Royle, 2020) frameworks offers promising avenues for integrating different classification systems and modelling classification errors.

AUTHOR CONTRIBUTIONS

Simone Santoro, Javier Calzada, Manuel Emilio Gegúndez and Santiago Gutiérrez-Zapata conceived the ideas, designed the methodology and collected the data. Simone Santoro, Manuel Emilio Gegúndez and Santiago Gutiérrez-Zapata analysed the data. Simone Santoro led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

We thank Zooniverse contributors, especially moderators, for supporting the Iberian Camera Trap Project. We also thank the two anonymous reviewers for their valuable feedback. This research was funded by Biodiversa+, the European Biodiversity Partnership, in the context of the WildINTEL project under the 2022–2023 BiodivMon joint call. It was co-funded by the European Commission (GA No. 101052342) and the following funding organisations: Agencia Estatal de Investigación (Spain, PCI2023-145963-2), National Science Centre (Poland, 2023/05/Y/NZ8/00104), the Research Council of Norway (Norway) and the German Research Foundation (Germany). Additional support was provided by Fundación Biodiversidad, Ministerio para la Transición Ecológica y el Reto Demográfico, Project AI-CENSUS and the Programa Operativo FEDER Andalucía 2014–2020, Project UHU-202028. SGZ acknowledges support from Santander SHEs grants at the University of Huelva. NS acknowledges support from a María Zambrano scholarship from the University of Huelva. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70132.

DATA AVAILABILITY STATEMENT

All data and code used for analyses in this manuscript are publicly available on GitHub: https://github.com/simonesantoro77/paperMEE25_supporting_information and Zenodo via https://doi.org/10.5281/zenodo.15785222 (Santoro et al., 2025).

ORCID

Simone Santoro https://orcid.org/0000-0003-0986-3278
Kate Brandis https://orcid.org/0000-0001-6807-0142
Simone Tenan https://orcid.org/0000-0001-5055-9193

REFERENCES

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). Springer International Publishing. https://beerys.github.io/CaltechCameraTraps/

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., & Stuart-Smith, J. F. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalised linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Burghardt, G. M., Bartmess-LeVasseur, J. N., Browning, S. A., Morrison, K. E., Stec, C. L., Zachau, C. E., & Freeberg, T. M. (2012). Perspectives–minimising observer bias in behavioral studies: A review and recommendations. *Ethology*, *118*(6), 511–517.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer-Verlag.

Burton, A. C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J. T., Bayne, E., & Boutin, S. (2015). Wildlife camera trapping: A review and recommendations for linking surveys to ecological processes.

Journal of Applied Ecology, *52*(3), 675–685. https://doi.org/10.1111/1365-2664.12432

Caravaggi, A., Banks, P. B., Burton, A. C., Finlay, C. M. V., Haswell, P. M., Hayward, M. W., Rowcliffe, J. M., & Wood, M. D. (2017). A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation*, *3*(3), 109–122. https://doi.org/10.1002/rse2.48

Chambert, T., Miller, D. A., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*(2), 332–339.

Chicco, D., & Jurman, G. (2022). An invitation to greater use of Matthews correlation coefficient in robotics and artificial intelligence. *Frontiers in Robotics and AI*, *9*, 876814. https://doi.org/10.3389/frobt.2022.876814

Cowans, A., Lambin, X., Hare, D., & Sutherland, C. (2024). Improving the integration of artificial intelligence into existing ecological inference workflows. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.14485

Cusack, J. J., Swanson, A., Coulson, T., Packer, C., Carbone, C., Dickman, A. J., Kosmala, M., Lintott, C., & Rowcliffe, J. M. (2015). Applying a random encounter model to estimate lion density from camera traps in Serengeti National Park, Tanzania: Density estimation of Serengeti lions. *The Journal of Wildlife Management*, *79*(6), 1014–1021. https://doi.org/10.1002/jwmg.902

Delisle, Z. J., Flaherty, E. A., Nobbe, M. R., Wzientek, C. M., & Swihart, R. K. (2021). Next-generation camera trapping: Systematic review of historic trends suggests keys to expanded research applications in ecology and conservation. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/10.3389/fevo.2021.617996

Devarajan, K., Morelli, T. L., & Tenan, S. (2020). Multi-species occupancy models: Review, roadmap, and recommendations. *Ecography*, *43*, 1612–1624. https://doi.org/10.1111/ecog.04957

Dussert, G., Chamaillé-Jammes, S., Dray, S., & Miele, V. (2024). Being confident in confidence scores: Calibration in deep learning models for camera trap image sequences. *Remote Sensing in Ecology and Conservation*. https://doi.org/10.1002/rse2.412

Efron, B., & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, *253*(5018), 390–395.

Fiske, I., & Chandler, R. (2011). Unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, *43*(10), 1–23. https://doi.org/10.1016/j.infsof.2008.09.005

Fortson, L., Masters, K., Nichol, R., Edmondson, E., Lintott, C., Raddick, J., & Wallin, J. (2012). Galaxy zoo. *Advances in Machine Learning and Data Mining for Astronomy*, *2012*, 213–236.

Fox, J., & Weisberg, S. (2019). *An {R} companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gaynor, K. M., Hojnowski, C. E., Carter, N. H., & Brashares, J. S. (2018). The influence of human disturbance on wildlife nocturnality. *Science*, *360*(6394), 1232–1235.

Gibbon, G. E., Bindemann, M., & Roberts, D. L. (2015). Factors affecting the identification of individual mountain bongo antelope. *PeerJ*, *3*, e1303.

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

Green, S. E., Rees, J. P., Stephens, P. A., Hill, R. A., & Giordano, A. J. (2020). Innovations in camera trapping technology and approaches: The integration of citizen science and artificial intelligence. *Animals*, *10*(1), 132. https://doi.org/10.3390/ani10010132

Hartig, F. (2021). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed). R package version 0.4.1 regression models*. Comprehensive R Archive Network (CRAN).

Hofmeester, T. R., Cromsigt, J. P. G. M., Odden, J., Andrén, H., Kindberg, J., & Linnell, J. D. C. (2019). Framing pictures: A conceptual framework to identify and correct for biases in detection probability of camera traps enabling multi-species comparison. *Ecology and Evolution*, *9*(4), 2320–2336. https://doi.org/10.1002/ece3.4878

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, *3*(4), 539–551. https://doi.org/10.1038/s41559-019-0826-1

Kellner, K. F., Smith, A. D., Royle, J. A., Kéry, M., Belant, J. L., & Chandler, R. B. (2023). The unmarked R package: Twelve years of advances in occurrence and abundance modelling in ecology. *Methods in Ecology and Evolution*, *14*(6), 1408–1415. https://doi.org/10.1111/2041-210X.14123

Kery, M., & Royle, J. A. (2020). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS: Vol. 2: Dynamic and advanced models*. Academic Press.

Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernández, N., García, E. A., Guralnick, R. P., Isaac, N. J. B., Kelling, S., Los, W., McRae, L., Mihoub, J. B., Obst, M., Santamaria, M., Skidmore, A. K., Williams, K. J., Agosti, D., Amariles, D., Arvanitidis, C., … Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews*, *93*(1), 600–625. https://doi.org/10.1111/brv.12359

Lenth, R. (2022). *emmeans: Estimated marginal means, aka least-squares means. R package Version 1.7.2*. Comprehensive R Archive Network (CRAN).

Lonsinger, R. C., Dart, M. M., Larsen, R. T., & Knight, R. N. (2024). Efficacy of machine learning image classification for automated occupancy-based monitoring. *Remote Sensing in Ecology and Conservation*, *10*(1), 56–71. https://doi.org/10.1002/rse2.356

Mackenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., & Hines, J. E. (2017). *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence*. Elsevier.

McClure, E. C., Sievers, M., Brown, C. J., Buelow, C. A., Ditria, E. M., Hayes, M. A., Pearson, R. M., Tulloch, V. J. D., Unsworth, R. K. F., & Connolly, R. M. (2020). Artificial intelligence meets citizen science to supercharge ecological monitoring. *Patterns*, *1*(7), 100109. https://doi.org/10.1016/j.patter.2020.100109

Meek, P., Ballard, G., Fleming, P., & Falzon, G. (2016). Are we getting the full picture? Animal responses to camera traps and implications for predator studies. *Ecology and Evolution*, *6*(10), 3216–3225. https://doi.org/10.1002/ece3.2111

Meek, P. D., Ballard, G., Claridge, A., Kays, R., Moseby, K., O'Brien, T., O'Connell, A., Sanderson, J., Swann, D. E., Tobler, M., & Townsend, S. (2014). Recommended guiding principles for reporting on camera trapping research. *Biodiversity and Conservation*, *23*(9), 2321–2343. https://doi.org/10.1007/s10531-014-0712-8

Monchy, C., Etienne, M.-P., & Gimenez, O. (2025). Using informative priors to account for identifiability issues in occupancy models with identification errors. *Peer Community Journal*, *5*. https://doi.org/10.24072/pcjournal.511

Norman, D. L., Bischoff, P. H., Wearn, O. R., Ewers, R. M., Rowcliffe, J. M., Evans, B., Sethi, S., Chapman, P. M., & Freeman, R. (2023). Can CNN-based species classification generalise across variation in habitat within a camera trap survey? *Methods in Ecology and Evolution*, *14*(1), 242–251. https://doi.org/10.1111/2041-210X.14031

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), 150–161. https://doi.org/10.1111/2041-210X.13504

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

O'Connell, A. F., Nichols, J. D., & Karanth, K. U. (Eds.). (2011). *Camera traps in animal ecology: Methods and analyses*. Springer.

Pradel, R. (2005). Multievent: An extension of multistate capture- recapture models to uncertain states. *Biometrics*, *61*(6), 442–447. https://doi.org/10.1111/j.1541-0420.2005.00318.x

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rhinehart, T. A., Turek, D., & Kitzes, J. (2022). A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution*, *13*(8), 1778–1789. https://doi.org/10.1111/2041-210X.13905

Rigoudy, N., Dussert, G., Benyoub, A., Besnard, A., Birck, C., Boyer, J., Bollet, Y., Bunz, Y., Caussimont, G., Chetouane, E., Carriburu, J. C., Cornette, P., Delestrade, A., De Backer, N., Dispan, L., Le Barh, M., Duhayer, J., Elder, J. F., Fanjul, J. B., … Chamaillé-Jammes, S. (2023). The DeepFaune initiative: A collaborative effort towards the automatic identification of European fauna in camera trap images. *European Journal of Wildlife Research*, *69*(6), 113. https://doi.org/10.1007/s10344-023-01742-7

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

Royle, J. A., & Link, W. A. (2006). Generalised site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*(4), 835–841.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Santoro, S., Gutiérrez-Zapata, S., Calzada, J., Selva, N., Marín-Santos, D., Beery, S., Brandis, K., Fernández de Viana, I., Meek, P., Mortelliti, A., Revilla, E., Rodríguez, J. P., Straková, L., Tenan, S., & Gegúndez, M. E. (2025). Data and code from: "Essential tools but overlooked bias: Artificial intelligence and citizen science classification affect camera trap data". Zenodo Digital Repository. https://doi.org/10.5281/zenodo.15785222

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 1–48.

Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., Smith, K., Revaz, B., Finnbogason, B., Szantner, A., & Lundberg, E. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature Biotechnology*, *36*(9), 820–832. https://doi.org/10.1038/nbt.4225

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

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

Torrey, L., & Shavlik, J. (2010). Transfer learning. In E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, & A. J. Serrano López (Eds.), *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques* (pp. 242–264). IGI Global.

Townsend, P. A., Clare, J. D. J., Liu, N., Stenglein, J. L., Anhalt-Depies, C., Van Deelen, T. R., Gilbert, N. A., Singh, A., Martin, K. J., & Zuckerberg, B. (2021). Snapshot Wisconsin: Networking community scientists and remote sensing to improve ecological monitoring and management. *Ecological Applications*, *31*(8), e02436. https://doi.org/10.1002/eap.2436

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M. C., Stewart, C. V., & Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, *13*(1), 792. https://doi.org/10.1038/s41467-022-27980-y

Urbanek, R. E., Ferreira, H. J., Olfenbuttel, C., Dukes, C. G., & Albers, G. (2019). See what you've been missing: An assessment of Reconyx® PC900 Hyperfire cameras. *Wildlife Society Bulletin*, *43*(4), 630–638. https://doi.org/10.1002/wsb.1015

Ushey, K., & Wickham, H. (2025). *renv: Project environments*. R package Version 1.1.4. https://cran.r-project.org/package=renv

Vélez, J., McShea, W., Shamon, H., Castiblanco-Camacho, P. J., Tabak, M. A., Chalmers, C., Fergus, P., & Fieberg, J. (2022). An evaluation of platforms for processing camera trap data using artificial intelligence. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.14044

Welbourne, D. J., Claridge, A. W., Paull, D. J., & Lambert, A. (2016). How do passive infrared triggered camera traps operate and why does it matter? Breaking down common misconceptions. *Remote Sensing in Ecology and Conservation*, *2*(2), 77–83. https://doi.org/10.1002/rse2.20

Whytock, R. C., Świeżewski, J., Zwerts, J. A., Bara-Słupski, T., Koumba Pambo, A. F., Rogala, M., Bahaa-el-din, L., Boekee, K., Brittain, S., Cardoso, A. W., Henschel, P., Lehmann, D., Momboua, B., Kiebou Opepa, C., Orbell, C., Pitman, R. T., Robinson, H. S., & Abernethy, K. A. (2021). Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, *2021*, 1–13. https://doi.org/10.1111/2041-210X.13576

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

SUPPORTING INFORMATION

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

Figure S1. Predicted versus true occupancy (psi) under psi = 0.2 and $p = 0.2$.

Figure S2. Predicted versus true occupancy (psi) under psi = 0.2 and $p = 0.5$.

Figure S3. Predicted versus true occupancy (psi) under psi = 0.2 and $p = 0.8$.

Figure S4. Predicted versus true occupancy (psi) under psi = 0.5 and $p = 0.2$.

Figure S5. Predicted versus true occupancy (psi) under psi = 0.5 and $p = 0.5$.

Figure S6. Predicted versus true occupancy (psi) under psi = 0.5 and $p = 0.8$.

Figure S7. Predicted versus true occupancy (psi) under psi = 0.8 and $p = 0.2$.

Figure S8. Predicted versus true occupancy (psi) under psi = 0.8 and $p = 0.5$.

Figure S9. Predicted versus true occupancy (psi) under psi = 0.8 and $p = 0.8$.

Figure S10. Main effect of sensitivity (recall) on RMSE of predicted occupancy.

Figure S11. Interaction between detection probability (p) and precision on RMSE of predicted occupancy.

Figure S12. Interaction between occupancy probability (psi) and detection probability (p) on RMSE of predicted occupancy.

Figure S13. Interaction between occupancy probability (psi) and precision on RMSE of predicted occupancy.

Figure S14. Interaction between occupancy probability (psi) and sensitivity (recall) on RMSE of predicted occupancy.

Figure S15. Main effect of detection probability (p) on RMSE of predicted occupancy.

Figure S16. Main effect of precision on RMSE of predicted occupancy.

Figure S17. Main effect of occupancy probability (psi) on RMSE of predicted occupancy.

Appendix S1. Confusion matrices for classifier performance.

Appendix S2. GLM summaries of differences in recall and precision.

Appendix S3. Simulation study: effects of classification errors on occupancy estimates.

How to cite this article: Santoro, S., Gutiérrez-Zapata, S., Calzada, J., Selva, N., Marín-Santos, D., Beery, S., Brandis, K., Fernández de Viana, I., Meek, P., Mortelliti, A., Revilla, E., Rodríguez, J. P., Straková, L., Tenan, S., & Gegúndez, M. E. (2025). Essential tools but overlooked bias: Artificial intelligence and citizen science classification affect camera trap data. *Methods in Ecology and Evolution*, *00*, 1–15. https://doi.org/10.1111/2041-210X.70132