The DeepFaune initiative: a collaborative effort towards the automatic identification of the French fauna in camera-trap images

Noa Rigoudy[1,2], Abdelbaki Benyoub[3], Aurélien Besnard[1], Carole Birck[4], Yoann Bollet[5], Yoann Bunz[3], Nina De Backer[6], Gérard Caussimont[7], Anne Delestrade[8], Lucie Dispan[9], Jean-François Elder[10], Jean-Baptiste Fanjul[5,11], Jocelyn Fonderflick[12], Mathieu Garel[13], William Gaudry[13], Agathe Gérard[14], Olivier Gimenez[1], Arzhela Hemery[1], Audrey Hemon[15], Jean-Michel Jullien[16], Maden Le Barh[9], Isabelle Malafosse[12], Malory Randon[17], Romain Ribière[18], Sandrine Ruette[13], Guillaume Terpereau[19,20], Wilfried Thuiller[19,20], Valentin Vautrain[6], Bruno Spataro[2], Vincent Miele[2,*], and Simon Chamaillé-Jammes[1,*]

[1]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France
[2]Université Lyon 1; CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Villeurbanne, France
[3]Parc National des Écrins, Domaine de Charance, Gap, France
[4]Asters, Conservatoire d'espaces naturels de Haute-Savoie, 84 route du Viéran, PAE de Pré Mairy, Prin, France
[5]Fédération Départementale des Chasseurs de l'Ain, 19 Rue du 4 Septembre, Bourg-en-Bresse, France
[6]Parc naturel régional du Perche, Courboyer, Perche-en-Nocé, France
[7]Fond d'Intervention Eco-Pastoral (FIEP) Groupe Ours Pyrénées, 1 rue Boyrie, Pau, France
[8]Centre de Recherches sur les Écosystèmes d'Altitude (CREA Mont-Blanc), Observatoire du Mont-Blanc, Chamonix, France
[9]CERFE - URCA, 5 rue de la Héronnière, Boult-aux-Bois, France
[10]Parc naturel régional des Marais du Cotentin et du Bessin, 3 village Ponts d'Ouve, Saint-Côme-du-Mont, Carentan-les-Marais, France
[11]Fédération Départementale des Chasseurs du Jura, Route de la Fontaine Salée, Arlay, France
[12]Parc national des Cévennes, 6 bis, Place du palais, 48400 Florac-Trois-Rivières, France
[13]Office Français de la Biodiversité, Direction de la Recherche et de l'Appui Scientifique, Montfort 01330 Birieux & 5 allée Bethleem Gières, France
[14]Parc naturel régional des Ballons des Vosges, gestionnaire de la RNN de la Tourbière de Machais, 1 place des verriers, Wildenstein, France
[15]Établissement public du Mont-Saint-Michel, 16 route de la Caserne, Beauvoir, France
[16]Chemin de Pré Rond-Les Mariages, Le Châtelard, France
[17]Fédération Départementale des Chasseurs de la Drôme, 3132 Route des Sétérées, Crest, France
[18]Parc naturel régional de l'Aubrac, Place d'Aubrac, Aubrac, France
[19]Université Grenoble Alpes, CNRS, Université Savoie Mont Blanc, LECA, Laboratoire d'Écologie Alpine, Grenoble, France
[20]Observatoire spatio-temporel de la biodiversite et du fonctionnement des socio-ecosystemes de montagne (ORCHAMP), France,

[*]Corresponding authors: vincent.miele@univ-lyon1.fr; simon.chamaille@cefe.cnrs.fr

Abstract

Camera-traps have revolutionized the way ecologists monitor biodiversity and population abundances. Their full potential is however only realized when the hundreds of thousands of images collected can be rapidly classified with minimal human intervention. Machine learning approaches, and in particular deep learning methods, have allowed extraordinary progress towards this end. Trained classification models remain rare however, and for instance are only emerging for the European fauna. This can be explained by the technical expertise they require but also by the limited availability of large datasets of annotated pictures, which are key to obtaining successful recognition models.

In this context, we set-up the DeepFaune initiative (`https://deepfaune.cnrs.fr`), a large-scale collaboration between dozens of partners involved in research, conservation and management of wildlife in France. The aim of DeepFaune is to aggregate individual datasets of annotated pictures to train species classification models based on convolutional neural networks, an established deep-learning approach.

Here we report on our first milestone, a two-step pipeline built upon the MegaDetector algorithm for detection (discarding empty pictures and cropping the animal) and a classification model for 18 species or higher-level taxa as well as people and vehicles. The classification model achieved 92% validation accuracy and showed $> 90\%$ sensitivity and specificity for many classes. Most importantly, these performances were generally conserved when tested on an independent out-of-sample dataset. In addition, we developed a cross-platform graphical-user-interface that allows running the pipeline on images stored locally on a personal computer.

In conclusion, the DeepFaune initiative provides a freely available (for non-commercial purposes) toolbox with high performance to classify the French fauna in camera-trap images.

1 INTRODUCTION

Camera-traps have revolutionized the way ecologists monitor biodiversity and population abundances (O'Connell et al., 2011; Howe et al., 2017). Relatively cheap, easy to deploy and autonomous, camera-traps enable to scale-up monitoring efforts dramatically, both in space and time. The continuous monitoring they provide also facilitates the detection of rare species. Therefore, and unsurprisingly, deployments of tens to hundreds of camera-traps are now common.

The full potential of camera-traps is however only realized when the hundreds of thousands of images, many being empty from spurious detections, can be rapidly classified with minimal human intervention (Chen et al., 2014; Schneider et al., 2019; Wearn et al., 2019; Tuia et al., 2021). Since the beginning, machine learning approaches, and in particular deep learning methods, have held the promise to solve this issue. Recent works have confirmed their power: for instance, deep-learning models developed by different teams (Tabak et al., 2019; Willi et al., 2019; Whytock et al., 2021) obtained $> 90\%$ recall and precision for a number of mammal species of North American ecosystems, African savannas and tropical forests.

The number of image classification models available to ecologists currently remains low, and their taxonomic coverage is still limited. So far, most of the image classification models developed on camera-trap data have been trained, even when based on millions of pictures, using images from one or two sites collected by a few partners. For instance, Norouzzadeh and colleagues (Norouzzadeh et al., 2018) trained their model on 3.2 million images taken from the Serengeti Snapshot project (Swanson et al., 2015) to identify 48 African mammals, yet all data was collected from one large study area in Serengeti National Park as part of the same monitoring project. Although this can lead to highly accurate models, when tested on data collected in the same sites, and even provide models which retain sufficient accuracy on out-of-sample data to be useful for others working at other sites, such approaches may not always be possible or desirable. For instance, we are not aware

of any European research group that has either collected such a large camera-trap image dataset or trained an equally-powerful image classification model on European fauna up to this day. We believe there is potential in aggregating multiple small datasets from a large number of partners to obtain sufficiently large training, validation and test datasets to develop efficient classification models.

The pros and cons of a cross-partner image aggregation strategy, and its ability to lead to successful models, has yet to be investigated. Recent initiatives such as Wildlife Insights, an online platform to manage and classify camera-trap images (`https://www.wildlifeinsights.org`), have embraced this strategy. Naturally, each initiative will have its own specific design choices which may represent drawbacks for certain institutional partners. For example, online platforms require the users to upload their images, and sometimes have non-optional data-sharing policies, an approach which may not be suitable for certain institutional partners which have legally-bounding contracts on data collection or privacy concerns.

We report here on the DeepFaune initiative (`https://deepfaune.cnrs.fr`). This initiative aims at (1) aggregating camera-trap images from many French institutions or research groups to create a common and large-scale dataset, made available to the whole community when possible, (2) developing a deep-learning based species classification pipeline that can also identify empty pictures, using a two-step approach based on running MegaDetector (Beery et al., 2019) to detect and crop animals in images followed by predicting from a convolutional neural network (CNN) model trained to classify the cropped images using the database built in (1), and (3) providing a free graphical-user-interface (GUI) so that non-computer savvy partners can run the classification model on self-hosted images with standard personal computers.

2 METHODS

2.1 Partners and data collection

The initiative brings together 38 partners (see the complete list at `https://deepfaune.cnrs.fr`), mostly composed of institutions managing protected areas, hunting federations or academic research groups across hexagonal France (Fig.1). Of the thirty-eight interested, thirty provided camera-trap images or videos, originating from various monitoring projects (e.g. general monitoring of Alpine fauna, monitoring of the burrows of badgers) or opportunistic surveillance (e.g. to assess the presence of wolves). One additional partner joined the project, after initial data collection, providing annotated pictures to use as an *out-of-sample* dataset for a stringent test of the model's accuracy (see below).

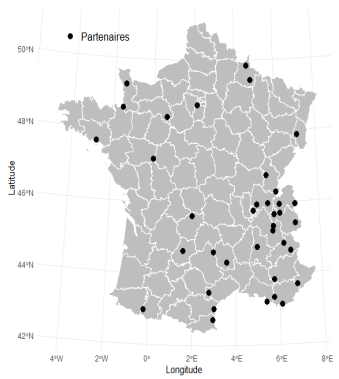


Figure 1: Distribution of the DeepFaune partners in France

2.2 Dataset

After removing corrupted or not fully annotated files, we gathered and sorted 422,765 annotated pictures and 4,460 annotated videos. These included either one of the 134 animal species identified, a person, a vehicle, or were empty, i.e. had no animal, person or vehicle visible. In some instances, annotation was made at a higher taxonomic level than the species level (e.g. bird, rodent). The number of medias provided by each partner varied a lot, from 20 videos to more than 140,000 images. As we wanted to train our model for camera-trap *images*, we converted videos into images by retaining one frame per second out of the 5 first seconds of each video.

We restricted our study to the identification of vehicles, people and 18 animal classes that were common enough in the images we had gathered: badger, bird, chamois (alpine and pyrenean), cow, dog, felinae (domestic/wild cat and lynx), red fox, ibex (alpine only), lagomorph, micromammal, mustelidae, mouflon, red deer, roe deer, sheep, squirrel, wild boar and wolf. We assumed that an image could contain only one of these classes. After keeping the relevant images, our final dataset was composed of 352,751 annotated images, including 2.1 % of empty images.

Our out-of-sample dataset originated from 27 camera traps installed in 11 different locations distributed along two altitudinal gradients in the Alps, France. It was made of 25,836 annotated pictures, containing images of humans (but no vehicles) and 16 identified animal species (felinae and mouflon were lacking), and 34.5% of empty images.

2.3 DeepFaune identification pipeline

To avoid shortcut learning, we chose to rely on a two-step approach: (step 1) detecting animals, persons or vehicles in images and filtering out empty images, and (step 2) using a CNN model to identify the species detected in the image. This approach has been the leading approach encountered in successive iWildcam competitions (Beery et al., 2021).

2.3.1 Filtering empty images with MegaDetector

We used MegaDetector v4.1 `https://github.com/microsoft/CameraTraps/` (Beery et al., 2019) to detect animals, people or vehicles that were present in each image. MegaDetector v4.1 is based on the image detection model Faster-RCNN (Ren et al., 2016), which allowed us to produce one cropped image per animal that was detected (there were potentially multiple individuals in a given image) or per human/vehicle. Therefore, we were able to filter out empty images as a first step. We used a threshold of 0.9 to only retain the elements which were detected with the highest scores. After this step, we remained with a dataset of 270,611 cropped images. Indeed, we observed that some images were annotated as containing an animal but this animal was very hard to distinguish. This was observed on images containing small animals (e.g. rodents, birds) which were detected far from the camera trap but visible to expert human eyes. This was also the case when images were annotated by humans using consecutive images. A single image containing a small part of an animal can be annotated by a human observer given the other images preceding or following the focal image. As a consequence, MegaDetector (Beery et al., 2019) was sometimes unable to find an animal on all the images of a given sequence.

2.3.2 Training a CNN model to classify images amongst 18 animal species, people and vehicles

We developed a CNN model to classify images amongst the 20 classes retained in the dataset: the 18 animal species of interest, as well as people and vehicles. The model was trained on the images cropped during the previous stage using MegaDetector. Since our original images were annotated, it was possible to annotate any cropped image using the annotation of its parent image. For instance,

if MegaDetector detected two animals in an image annotated as 'wild boar', we considered that the two resulting cropped images contained a wild boar. The output of the CNN model is, for each image processed, a prediction score expressing the certainty of the attribution of the image to each class. A score of 1 in one class indicates that the model is certain that an object of this class is present in the image. Conversely, a score of 0 in one class indicates that the model is certain that an object of this class is not present in the image.

Building independent train and validation datasets. CNN models require independent training and validation datasets to learn successfully, and as commonly done we split the whole image dataset into a train dataset containing 90 % of the images, and a validation dataset containing the rest. One issue that could arise during this split is that pictures collected within the same *sequence* (i.e. a batch of pictures taken a few seconds apart) are used in both the train and validation datasets, which would threaten their independence. As we obtained pictures collected with many different camera-trap models, there was no consistency in the EXIF information that we could use, and in particular could not always obtain the date information. We chose to use the image filenames instead. We used a heuristic based on text mining in the R package `stringdist` to compare filenames to each other. We considered that two files were not independent if their filenames were too similar, i.e. if their similarity was above 0.9, which was conservative. This way, we were certain that there was no image from a given sequence or video that were present in both the train and the validation datasets.

Coping with class imbalance. Class imbalance, i.e. the fact that the number of images per class differs), is known to affect both training and validation (Johnson and Khoshgoftaar, 2019). Here we propose a novel approach that combines downsampling and upsampling. At each epoch, we downsampled the class which was overrepresented down to a multiple of the rarest class. We chose to have a ratio of 5 between the number of images of the most common class and the number of images of the rarest class. This was estimated for every epoch, such that every image is seen by the model if we fit the model for enough epochs. As a consequence, images of the rarest classes were used in many epochs and can thus be considered *upsampled* whereas those of the most common classes are *downsampled*. This way, the optimization problem was different at every epoch, but more balanced. We considered that the stochasticity induced by the sampling process at each epoch had a positive impact on the model fit procedure and we observed a continuous reduction in model loss through epochs.

We used the same idea to handle imbalance during validation, and used the same ratio of 5 to create a more balanced validation dataset, which was not changed among epochs. This enabled us to compute what we called a *balanced* validation accuracy.

Training with transfer learning and image augmentation. We used transfer learning using an EfficientNetB3 model (Tan and Le, 2019) that was pre-trained on Imagenet, with a resolution of 300×300, using TensorFlow-Keras. Additionally, we performed image augmentation using the `imgaug Python` library (`https://github.com/aleju/imgaug`). Each image in each batch was modified with a random set of transformations such as horizontal flip, affine transformations, gray scale transformation, blurring, brightness or contrast changes. We used the softmax output as prediction score. We estimated the model for a maximum of 120 epochs, monitored the overall balanced validation accuracy and stopped the estimation when it did not increase with a patience of 10 epochs. This procedure took about half a day with 8 CPUs and a Titan X GPU. We also report sensitivity (the true positive rate, i.e. the % of images with species i that are correctly classified as showing species i) and specificity (the true negative rate, i.e. the % of images without species i that are correctly classified as not showing species i).

2.4 Predictive performance on out-of-sample data using sequences

We investigated the ability of our two-step pipeline to perform accurate identification of species in an out-of-sample dataset, i.e. in images taken in contexts that have never been seen during the training stage (in other words, images from new camera traps deployed at new locations). The out-of-sample dataset contained pictures of most classes used in the training of the model, but for mouflon, felinae, cow and vehicles.

We first assessed the quality of the detection step based on Megadetector, using the rate of false negative (FN) detection – not detecting a species on an image where a human detected it – as metric. A high FN rate indicates that Megadetector commonly fails to detect the animal, person or vehicule in the picture. We compute the FN rate at the image and sequence level. A false negative sequence is one for which all images of the sequence are false negative.

We then explored the classification accuracy of the CNN-based classification step of the pipeline. We predicted the species present at the scale of the sequence, not at the scale of the individual image. Indeed, as commonly done by practitioners in the field, camera traps were set to take three successive images when triggered. Additionally, it is common that animals stay several seconds near the camera, triggering it multiple times in a row. Such events however require a single species identification. In this dataset, we could obtain the date and time of each image using the EXIF information. We considered that two images taken less than a few seconds apart, at the same site, could be considered to represent a *sequence* of images that captured the same event. Here we chose a threshold duration of 20s to consider that two images belongs to the same sequence. To obtain the classification at the sequence level, we first computed the prediction score for each image of the sequence that was not predicted as empty. We then averaged these scores and considered that the class with the highest average score was the predicted class for the sequence. Images that were not in a sequence were classified using their own prediction score.

2.5 The DeepFaune graphical user interface

In our pipeline, MegaDetector represented a bottleneck as it is very slow when used on a CPU (about 20 to 40 second per image in our experiments on various personal computers). We therefore developed a much faster alternative using Yolo v4 (Bochkovskiy et al., 2020), exploiting the cropped images used in this study as a training set for object detection. We then implemented the two-step approach developed in the present study using this alternative detector (about 1 second per image) and our CNN classifier inside a free graphical user interface available at `https://deepfaune.cnrs.fr`. This interface only requires the installation of Python v3 and its TensorFlow v2 module.

3 RESULTS

3.1 Performance of the CNN model on the validation dataset

We obtained an overall balanced validation accuracy of 92%. Sensitivities and specificities were generally good to very good, $> 90\%$ in many instances (Table 1). Among issues to be resolved in the future, we observed identification mismatches between the different ungulate species, mainly on images where only a part of the animal was present (for instance, the back of a roe deer was recognized as a red deer). Without surprise, the lower performance among ungulates was observed for ibex and mouflon for which we had less than a thousand images in the train dataset. The model performed very badly on micromammals, and we observed that they were often mistaken for mustelidae, suggesting that this effect could also partly explain the poor model performance for this class. Finally, the quality of the classification of dog and people was only medium, likely because many images containing one of this class also contained the other (e.g. hunters with dogs).

Class	Nb images in train	Nb images in validation	Sensitivity (%) in validation	Specificity (%) in validation
badger	5412	707	95	98
ibex	732	241	95	79
red deer	38739	4701	97	95
chamois	44210	4311	98	97
roe deer	24365	3215	95	95
dog	8045	1180	90	88
squirrel	3599	440	89	97
felinae	1472	283	93	96
human	12561	947	88	87
lagomorph	5086	676	91	97
wolf	4844	424	86	95
micromammal	953	11	39	53
mouflon	497	60	78	95
sheep	5121	510	97	98
mustelide	2316	377	75	67
bird	39819	4738	99	97
red fox	20437	2398	95	96
wild boar	18242	2559	97	97
cow	2020	297	95	99
vehicle	3622	342	81	97

Table 1: Classification performance metrics computed on the validation dataset

3.2 Performance of the whole pipeline on the out-of-sample dataset

During the detection stage based on MegaDetector, we observed that FN rate on individual images where generally near or below 10% for the largest species, but much higher for smaller species (Figure 2): the FN rate were above 25% for mustelidae, micromammals, and reached near 50% for birds. We note however that the magnitude of the difference differed between species, but was likely important enough for a number of them to justify using the sequence level since the detection stage.

The CNN-based classification step achieved a overall balanced validation accuracy of 92.5%, and an overall validation accuracy (i.e. for the whole unbalanced validation dataset) of 95%. The best performance was achieved for human, squirrel and for large ungulates, with specificity and sensitivity being greater than 90% for red deer, roe deer, chamois, and to a lesser extent for wild boar and ibex (Table 2). There however remains some mismatches between these species (see Appendix S1), with for instance 16 sequences of red deer being classified as roe deer. Surprisingly, performance remained good for micromammals and we did not encounter the low performance noted for this taxa with the validation dataset (see Table 1). Specificity for red fox was low, with ten and eight sequences of wild boar and red deer predicted as red fox, respectively. This was not expected, given the very high performance achieved for this species in the validation dataset. Finally, the performance for images containing a wolf was low, indicating that our model might be under-performing for this species. Manual inspection revealed that vegetation in the foreground could be the source of this issue (see Figure 4), which therefore might be specific to this dataset. Also, and unexpectedly, we observed very few confusions between wolf and dog, suggesting that our model could distinguish between the two.

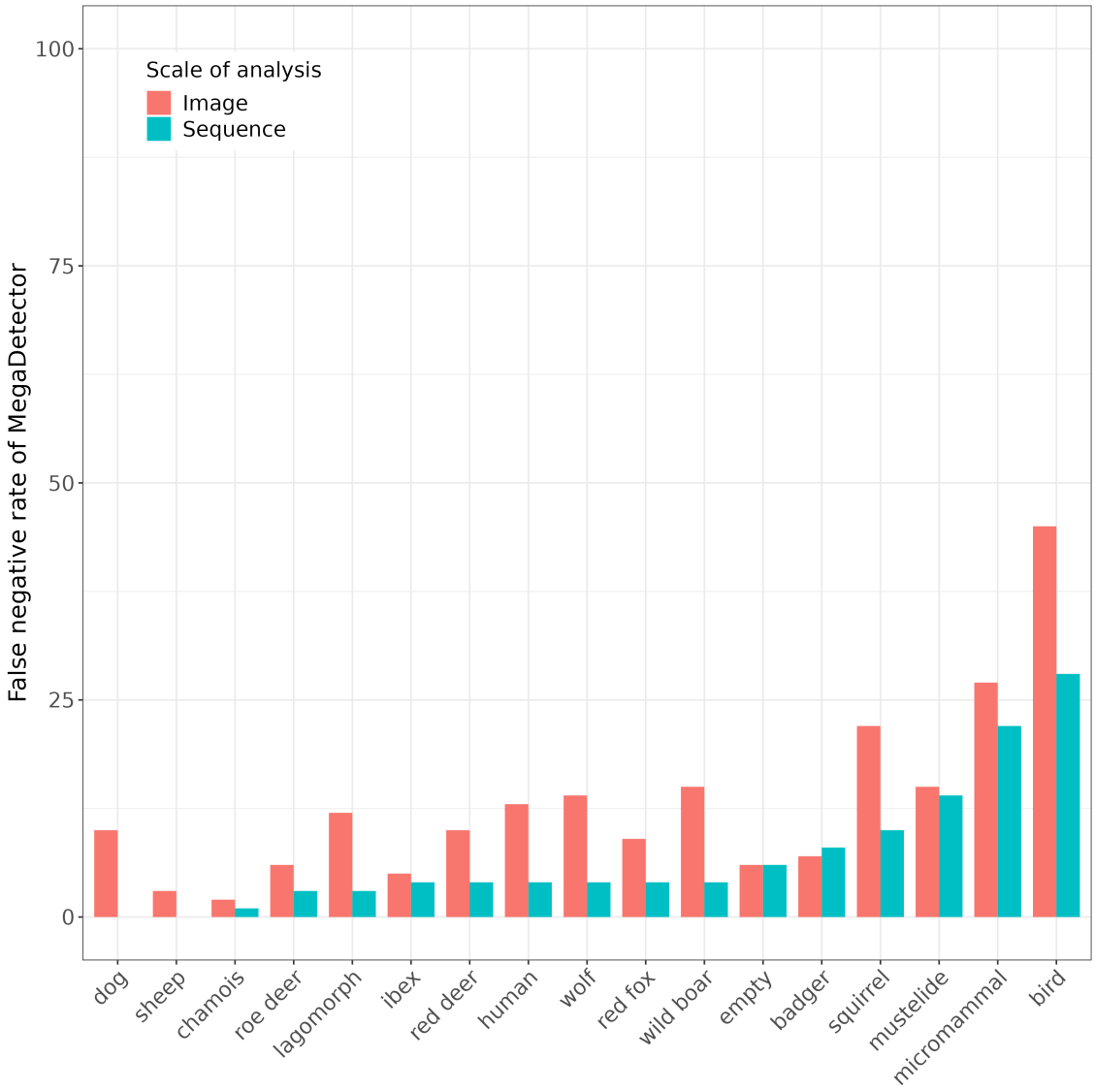


Figure 2: False negative rates obtained with MegaDetector on the out-of-sample dataset, computed at the scale of individual images or at the scale of image sequences. A false negative image is when the image of an animal was predicted as empty by MegaDetector. A false negative sequence is a sequence of images for which all the images are false negative. By construction, the false negative rate is necessarily lower or equal for sequences than for images.

4 DISCUSSION

The DeepFaune initiative represents a successful multi-partner collaboration to aggregate camera-trap images and build one of the first pipelines readily available to automatically classify camera-trap images collected in Europe. Our current model allows for predicting the presence of 20 different classes in camera-trap images (animal species or higher order taxa, as well as humans and vehicles). When used in conjunction with MegaDetector, it enables to analyze datasets in which empty images are numerous. Although we did use the standard approach of transfer learning, we implemented a number of tricks to deal with class imbalance and the independence of train and validation datasets that could be useful to others. Ultimately, our species classification model performed extremely well on the validation dataset and provided robust results on out-of-sample data. Additionally, we

Taxa	Sensitivity (%) in out-of-sample with sequences (or images*)	Specificity (%) in out-of-sample with sequences (or images*)
badger	100 (93)	60 (37)
ibex	95 (71)	96 (96)
red deer	79 (72)	92 (70)
chamois	90 (88)	96 (90)
roe deer	94 (88)	95 (88)
dog	100 (93)	48 (07)
squirrel	92 (83)	93 (79)
felinae	-	-
human	95 (92)	99 (98)
lagomorph	84 (72)	88 (55)
wolf	67 (62)	44 (29)
micromammal	85 (84)	96 (93)
mouflon	-	-
sheep	84 (70)	64 (96)
mustelide	84 (77)	53 (51)
bird	95 (96)	85 (44)
red fox	94 (88)	59 (39)
wild boar	88 (84)	89 (64)
cow	-	-
vehicle	-	-

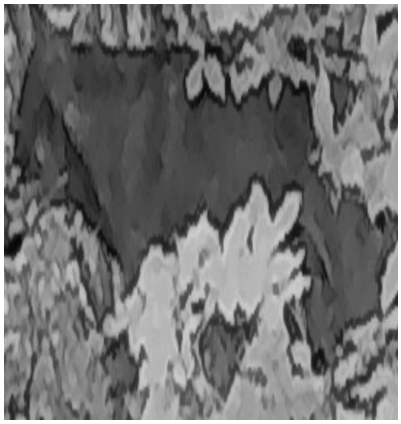
Table 2: Classification performance metrics computed on the out-of-sample dataset, using images or sequences. *Assessment of model performance at the image level is here not fair to the model, as people who annotated the images had the complete image sequence in hand and made deductions.

(a) Predicted as roe deer (0.97) (b) Predicted as roe deer (0.91) (c) Predicted as dog (0.75)

Figure 3: Using sequences decrease the false positive rate. A sequence of three images of a roe deer, taken over a 10-seconds interval (and cropped by MegaDetector). The third image is misclassified as a dog, but the average score at the sequence scale is highest for roe deer.

built a graphical user interface (GUI) so that partners can run the model locally on regular personal computers and automatically sort their newly acquired images based on the model predictions.

(a) Predicted as roe deer (score 0.67) in a sequence of 1 image



(b) Predicted as bird (score 0.93) 4 times in a sequence of 5 images

(c) Predicted as red deer (score 0.75) in a sequence of 1 image

Figure 4: Examples of sequences where the animal is not correctly identified as a wolf by our pipeline, possibly due to vegetation in the foreground.

4.1 Model performance and relevance for ecological studies

There are currently very few species-recognition models available for European fauna, and our results could therefore be used to benchmark future works.

Even by deep-learning classification standards, the quality of our model appears very good and compares favorably with results from similar exercises conducted on fauna from other continents. For instance, overall accuracy of 97%, 78% and 94% on validation datasets were respectively reported in studies on North American fauna (Tabak et al., 2019), Central African fauna (Whytock et al., 2021) and East African fauna (Norouzzadeh et al., 2018). Our model also improves on the one reported by Carl and colleagues (Carl et al., 2020), which is one of the rare models focused on European fauna, as we achieved much higher accuracy at the species level.

We managed to obtain good results, despite the moderate size of our dataset by deep learning standards, by overcoming a number of obstacles. Firstly, the large imbalance (two orders of magnitude more images of chamois than ibex, for example) would have biased the classification towards the most common species. We therefore implemented an original approach combining downsampling and upsampling techniques that successfully prevented the emergence of a relationship between identification performance and the number of images per class. Secondly, deep learning models are known to be able to learn 'shortcuts' (Geirhos et al., 2020). In our context, this would correspond to learning contextual elements (e.g. snow, peculiar vegetation) that would be associated with the species, rather than learning to recognize the species itself. To avoid this potential issue, we chose to classify cropped images, and not whole images (as opposed to Whytock et al. (2021)), with an additional procedure of image augmentation.

Overall, class-specific sensitivities and specificities are high enough for the model to be useful for many specific studies, some of which we highlight now: (1) automatized monitoring of large ungulates, a guild of important management interest in Europe. Ungulates are indeed generally very well classified by the model, with sensitivity and specificity values above $> 90\%$ for most species (as explained in the Results section, performance for ibex and mouflon will likely improve as new pictures are added in the database). These results suggest that our model could be a useful tool to facilitate studies investigating the effects of management practices on locally abundant ungulates (e.g. can wild boar population dynamics be controlled by hunting?), or the dynamics of prey under predation (e.g. do roe deer populations decrease as wolves return?); (2) monitoring large

mammalian communities. Although generally our model did not obtain accuracy levels as high as for ungulates for the other classes, the model's performance and taxonomic coverage suggest that it could be useful to automatically sort out hundreds of thousands of pictures for a range of taxonomic groups. This could prove especially useful in studies looking at the impact of anthropization on large mammal communities for example. A human intervention would still be required to manually verify the pictures for which prediction scores are low or to identify the species present in images where the model classifies at a higher taxonomic level (e.g. mustelids), but this effort should be dramatically reduced when using our model; (3) quantifying human disturbance levels. Classes related to human activities (human, dog, vehicle) are generally not as well identified as others, as we had only a limited number of images for these classes, and with limited diversity. Getting more images of people from partners might be difficult given privacy-aware policies and legislation, so we aim to solve this issue in the future by mining authorized images of people online or using existing datasets. As MegaDetector now directly integrates the classification of people with good success (Fennell et al., 2021), we might rely on it in the future for this class. The current model performance should however be sufficient to already build reliable metrics estimating human activities along wide gradients.

We emphasize here that even though sensitivities and specificities might be generally high, caution should be exerted when using the model to classify them, as relevant pictures might have been discarded at the detection stage by MegaDetector. This is clearly demonstrated by our out-of-sample test on bird pictures, which can be well classified by the model (arguably at this broad taxonomic level) but will often be discarded at the detection stage as the size of the bird is too small for MegaDetector to identify its presence. Pictures of micromammals are also often mistakenly discarded by MegaDetector, but this is less problematic as the model performs poorly on these anyway, and therefore no false confidence should occur. Taxa that require particular attention are squirrels and mustelids.

Importantly, the general high quality of the results of our two-step pipeline (detection+classification) was conserved when applied to an out-of-sample dataset. It is a common error to take model results on validation datasets at face value and assume similar accuracies will be observed in new applications. Previous studies (Whytock et al., 2021) have shown that this is not the case, with often dramatic declines in accuracy, to the point where the usefulness of the model could sometimes be called into question. Our out-of-sample test, based on heterogeneous images from 11 locations spread over two altitudinal gradients in France, suggests that this is not the case here. Additionally, this out-of-sample test was designed to test the applicability of our pipeline to real-world scenarios. It revealed that (1) although MegaDetector is generally efficient, the rate of false negatives, i.e. when the image is assumed empty and thus discarded, can be significant ($> 25\%$) for small animals, micromammals and birds in particular. This is a critical point that should be kept in mind when designing studies and interpreting results with our model, and warrants specific attention and, ideally, study-specific quantifications; (2) using the average prediction score over temporally-close pictures (i.e. sequence) can improve classification results if classification at the sequence-level is sufficient. This arose because, over a few pictures, at least one image is likely to be classified very reliably by the model (e.g. prediction score > 0.95), increasing largely the average prediction score for this class, and successfully identifying the animal present in all images of the same sequence. Also, confusion between species are less likely to occur over several images. Sequences of pictures are commonly used in camera-trapping to account for differences in fields of detection, camera views, and generally to increase chances of getting at least one identifiable picture of the animal triggering the camera-trap. We demonstrate here that this approach can be leveraged when using classification models. Our work also highlights that, when using sequence-based identification, the performance of classification models trained *without* sequence information are likely to be underestimated in contrast to what can be achieved at the sequence level. In the future, sequence or temporal information could be directly integrated into the training step, as done in context CNN

models (Beery et al., 2020; Beery et al., 2021; Tuia et al., 2021). We could not use this approach here as many partners provided data without information about the specific trapping sites where the pictures came from. We therefore could not reconstruct the sequences in the training dataset from the EXIF time-stamp information.

4.2 Lessons learned from a successful multi-partner initiative

One of the main strengths of the DeepFaune initiative lies in the creation of a nation-wide network among key actors in French biodiversity research, conservation and management. Under the lead of an academic research group originating from two distinct laboratories, 30 partners have shared camera-trap pictures and videos allowing to build what is likely to be one of the largest databases of camera-trap medias in France, in both numbers (over 420,000 photos and 4,400 videos) and taxonomic coverage.

This success should not hide the technical challenges of working with such a large number of actors. Building the picture database clearly revealed the intricacies of dealing with multiple, high-volume data transfers as well as with the strong heterogeneity in data acquisition and organization among the partners. The devil is in the details, and harmonizing directory, file and species names were all very time-consuming tasks that had to be dealt with by combining automatic (e.g. bash scripts) and manual interventions. Partner-specific data management appears as one of the strongest barriers to creating efficient large-scale datasets that do not depend on a single monitoring program (as opposed to the SerengetiSnapshot dataset for instance). Interestingly however, some of our individual partners were actually members of the same institution: for example, we received data from several national parks, which belong to the same institution (Parc Nationaux de France). Similarly, several teams from the Office National de la Biodiversité shared data, all having different formats. In such case, it would seem beneficial for the master institution to provide detailed data management guidelines that would allow standardized data management internally. In this context, centralized data management platforms that enforce data standardization, developed either within institutions or at national or international levels, might facilitate future works.

Having numerous and diverse partners is however key to ground our work in the reality of end-users, as the DeepFaune initiative was originally conceived to develop an easy-to-access tool of sufficient quality for field practitioners. Beyond the data collection stage, a regular communication between the leading academic team and the partners was critical to identify key expectations and potential difficulties in appropriation. Expectations shared by all partners were two-fold: (1) all partners expected a high performance of the classification of empty pictures for the pipeline to be useful, as these usually represent a large share of collected pictures. In the current version of the pipeline, we rely on MegaDetector and are thus bound to its own performance, which are generally good; (2) the pipeline should be easy to implement and run on a standard personal computer. This is long-standing issue in the distribution of deep-learning models, as even when training is not required and only predictions expected, the installation of libraries necessary for computation (e.g. Tensorflow) can be difficult. Uploading pictures on an online platform running its own servers naturally solves this issue, but we found that many partners were not tempted by this approach, in particular because of the need to upload gigabytes of images online, and sometimes because of the data sharing policies that could be enforced by the platforms. In the face of this, we decided to develop a Python-based GUI running the model, which can easily be launched on both Windows and Linux computers. Feedback from the partners will be critical to ensure that our approach is a credible solution. The developments of the GUI can be followed at `https://plmlab.math.cnrs.fr/deepfaune/software`.

An aspect for which partners' expectations sometimes differ concerns the classification model and how we use its prediction to sort pictures. What makes an 'optimal' sorting of pictures might in fact vary between partners. Some partners are interested in minimizing false negatives over

false positives. This can be the case when studying sensitive species, such as wolves, as positive classifications will in any case be verified manually. For other studies, the balance between false positives and false negatives may matter less. This is generally the case in occupancy modeling studies of common species (Gimenez et al., 2021). We also foresee that prediction scores will, in a near future, be directly used in the inference process allowing the uncertainty of the classification to be propagated into the estimation of any metric of interest. Irrespective of whether this will be successful or not, in response to this diversity of needs, we decided to let the users decide of a threshold of the prediction score below which the images will not be classified and requires manual inspection. We believe that this approach gives an important level of flexibility and that users will be able to learn what threshold works best for them by trial-and-error.

4.3 Conclusion

In conclusion, our work provides a rather successful classification model of the European fauna in camera-trap images, and which can be easily used by practitioners on self-hosted images using a standard personal computer. Such an approach contrasts with some recent developments that favor cloud-computing. Our work however remains a work-in-progress and feedbacks on the use of the model and its GUI, as well as annotations of images for which the model failed, should allow to improve the toolbox further.

5 AUTHORS' CONTRIBUTIONS

N. Rigoudy set up the collaborative network with S. Chamaillé-Jammes, organized data collection, and cleaned-up and managed the dataset with V. Miele. V. Miele developed the model and the graphical-user-interface. S. Chamaillé-Jammes initiated the project with N. Rigoudy and now manage it with V. Miele. B. Spataro provided assistance with the use of the LBBE PRABI computing cluster. All four make up the DeepFaune team. Other authors contributed data on their behalf or on behalf of their institutions.

6 ACKNOWLEDGMENTS

This work was performed using the computing facilities of the CC LBBE/PRABI. Funding was provided by the French National Center for Scientific Research (CNRS). This work has benefited from discussions initiated in the Statistical Ecology Research Group (EcoStat) of the CNRS. V. Miele would like to thank the LECA laboratory for having hosted him in Chambéry, and L. Humblot for insightful discussions about the GUI. The DeepFaune team would like to thank E. Chetouane for his contribution to the GUI.

Also, for many partners, field technicians and students contributed largely to data collection, and they are acknowledged for their critical work.

Some partners could not share images at the time of our study but actively participated in the discussions around future model developments: PNR des Ardennes, PNR de Chartreuse, PNR du Haut-Languedoc, PNR du Haut Jura, PNR du massif des Bauges, PNR de Préalpes d'Azur, PN de la Vanoise and PN des Calanques.

The following funders supported data collection through various programs (supported partners in parenthesis): Region Auvergne-Rhône-Alpes (FDC Drôme, Ain, Jura; CREA), Réseau de Transport d'Electricité (CEFE), Fédération Nationale des Chasseurs (FDC Ain, Jura, Office Français de la Biodiversité (through the Ecocontribution, FDC Ain, Jura), Conseils Départementaux de l'Ain et du Jura (FDC Ain, Jura), Fonds européens FEDER/POIA (CREA), Département de la Haute Savoie (CREA), Communauté de Communes Vallée de Chamonix (CREA), DREAL Bourgogne-Franche-Comté (FDC Jura), DREAL Auverge-Rhône-Alpes (FDC Ain), DREAL Normandie (PNR

Perche), DREAL Centre-Val-de-Loire (PNR Perche), Fédéraction Départmentale des Chasseurs de la Marne (CERFE), Réseau SNCF (CERFE), Autoroutes SANEF (CERFE).

7 DATA AVAILABILITY

Not all partners wanted or could, because of legally-binding contracts, release their images. A large proportion of our image dataset will however be made available to the scientific community at `https://deepfaune.cnrs.fr`, without metadata.

References

Beery, S., Agarwal, A., Cole, E., and Birodkar, V. (2021). The iwildcam 2021 competition dataset. arXiv preprint arXiv:2105.03494.

Beery, S., Morris, D., and Yang, S. (2019). Efficient pipeline for camera trap image review. arXiv preprint arXiv:1907.06772.

Beery, S., Wu, G., Rathod, V., Votel, R., and Huang, J. (2020). Context r-cnn: Long term temporal context for per-camera object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13075–13085.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection.

Carl, C., Schönfeld, F., Profft, I., Klamm, A., and Landgraf, D. (2020). Automated detection of european wild mammal species in camera trap images with an existing and pre-trained computer vision model. European Journal of Wildlife Research, 66(4):1–7.

Chen, G., Han, T. X., He, Z., Kays, R., and Forrester, T. (2014). Deep convolutional neural network based species recognition for wild animal monitoring. In 2014 IEEE international conference on image processing (ICIP), pages 858–862. IEEE.

Fennell, M., Beirne, C., and Burton, A. C. (2021). Use of object detection in camera trap image identification: assessing a method to rapidly and accurately classify human and animal detections for research and application in recreation ecology. bioRxiv preprint bioRxiv:2022.01.14.476404.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11):665–673.

Gimenez, O., Kervellec, M., Fanjul, J.-B., Chaine, A., Marescot, L., Bollet, Y., and Duchamp, C. (2021). Trade-off between deep learning for species identification and inference about predator-prey co-occurrence: Reproducible r workflow integrating models in computer vision and ecological statistics. arXiv preprint arXiv:2108.11509.

Howe, E. J., Buckland, S. T., Després-Einspenner, M.-L., and Kühl, H. S. (2017). Distance sampling with camera traps. Methods in Ecology and Evolution, 8(11):1558–1565.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1):1–54.

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

O'Connell, A. F., Nichols, J. D., and Karanth, K. U. (2011). Camera traps in animal ecology: methods and analyses, volume 271. Springer.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks.

Schneider, S., Taylor, G. W., Linquist, S., and Kremer, S. C. (2019). Past, present and future approaches using computer vision for animal re-identification from camera trap data. Methods in Ecology and Evolution, 10(4):461–470.

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. Sci Data, 2:150026.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., et al. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. Methods in Ecology and Evolution, 10(4):585–590.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, pages 6105–6114. PMLR.

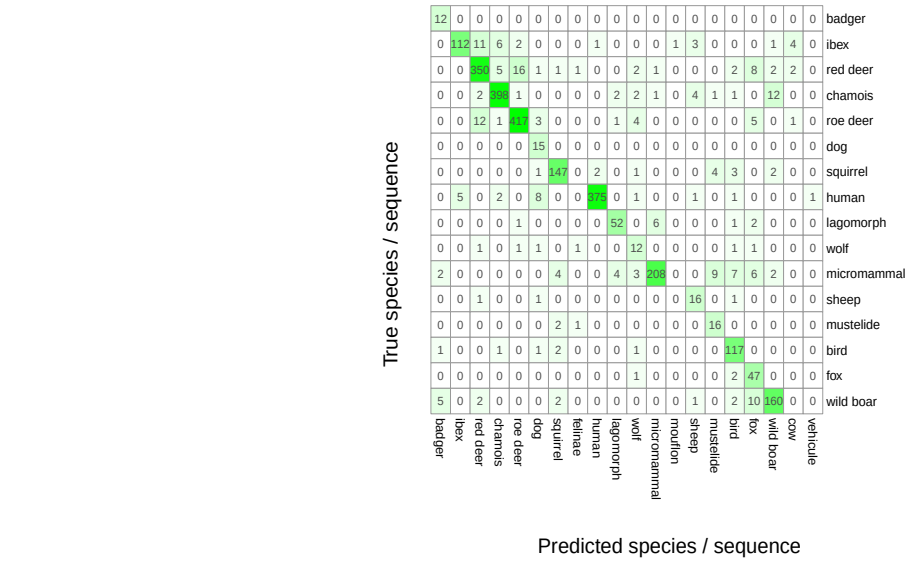
Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., et al. (2021). Seeing biodiversity: perspectives in machine learning for wildlife conservation. arXiv preprint arXiv:2110.12951.

Wearn, O. R., Freeman, R., and Jacoby, D. M. (2019). Responsible ai for conservation. Nature Machine Intelligence, 1(2):72–73.

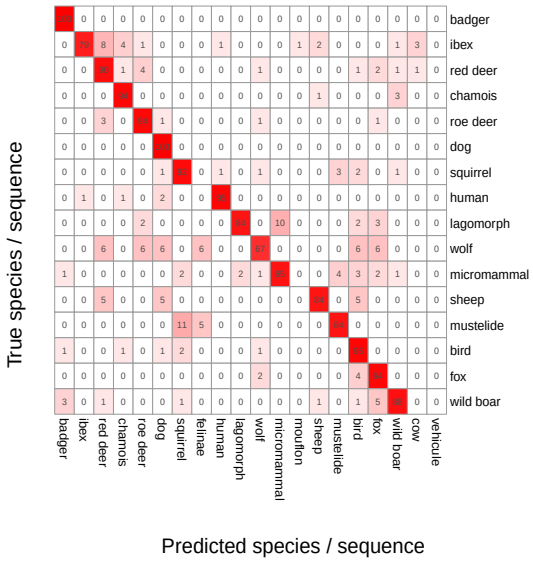
Whytock, R. C., Świeżewski, J., Zwerts, J. A., Bara-Słupski, T., Koumba Pambo, A. F., Rogala, M., Bahaa-el din, L., Boekee, K., Brittain, S., Cardoso, A. W., et al. (2021). Robust ecological analysis of camera trap data labelled by a machine learning model. Methods in Ecology and Evolution.

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., and 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.

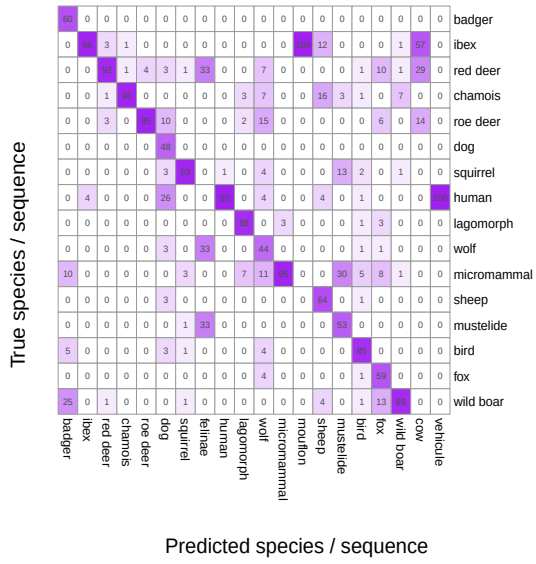
8 APPENDIX



(a) Raw number of sequences



(b) Sensitivity



(c) Specificity

Figure S1: Confusion matrices of the predictions of the CNN-based classifier on the out-of-sample dataset. Images were first cropped with MegaDetector.