Identification of morphologically cryptic species with computer vision models: wall lizards (Squamata: Lacertidae: *Podarcis*) as a case study

CATARINA PINHO[1,2,*], ANTIGONI KALIONTZOPOULOU[3], CARLOS A. FERREIRA[4] and JOÃO GAMA[4,5]

[1]*CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal*
[2]*BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal*
[3]*Department of Evolutionary Biology, Ecology and Environmental Sciences, and Biodiversity Research Institute (IRBio), Universitat de Barcelona, E-08028 Barcelona, Catalonia, Spain*
[4]*INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal*
[5]*FEP - University of Porto, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal*

Received 16 February 2022; revised 6 June 2022; accepted for publication 7 September 2022

Automated image classification is a thriving field of machine learning, and various successful applications dealing with biological images have recently emerged. In this work, we address the ability of these methods to identify species that are difficult to tell apart by humans due to their morphological similarity. We focus on distinguishing species of wall lizards, namely those belonging to the *Podarcis hispanicus* species complex, which constitutes a well-known example of cryptic morphological variation. We address two classification experiments: (1) assignment of images of the morphologically relatively distinct *P. bocagei* and *P. lusitanicus*; and (2) distinction between the overall more cryptic nine taxa that compose this complex. We used four datasets (two image perspectives and individuals of the two sexes) and three deep-learning models to address each problem. Our results suggest a high ability of the models to identify the correct species, especially when combining predictions from different perspectives and models (accuracy of 95.9% and 97.1% for females and males, respectively, in the two-class case; and of 91.2% to 93.5% for females and males, respectively, in the nine-class case). Overall, these results establish deep-learning models as an important tool for field identification and monitoring of cryptic species complexes, alleviating the burden of expert or genetic identification.

ADDITIONAL KEYWORDS: convolutional neural networks – cryptic species – deep learning – image classification – lizards.

INTRODUCTION

Despite the conceptual difficulties associated with the definition of species [reviewed in Zachos (2016)], naming and identifying species is a task of utmost pertinence for modern science, and taxonomy is considered a fundamentally important discipline (Wilson, 2004). Biological research requires a common system for classifying, naming and identifying species diversity that can be used across all disciplines and between different researchers. Moreover, species as biological units, and their corresponding names, are important well beyond the strict context of systematics, transcending numerous other fields of biology, such as ecology, evolution, biodiversity monitoring and conservation; and are also key to other scientific disciplines, such as medicine, pharmacology, agriculture and international trade legislation.

Current taxonomic research faces several important challenges: first, the acknowledgement that the biodiversity of Earth is far from completely described (and that the magnitude of this gap is

[*]Corresponding author. E-mail: catarina@cibio.up.pt

also hard to uncover; Costello *et al.*, 2013; Hortal *et al.*, 2015). Second, the fact that Earth is facing its sixth (and largely human-induced) mass extinction (Ceballos *et al.*, 2015) and that correctly cataloguing and identifying species is critical for monitoring and preserving biodiversity. Third, the well-known decrease of the taxonomic workforce over the past decades, a problem known as the 'taxonomic impediment' (Hopkins & Freckleton, 2002; Engel *et al.*, 2021). Together, these challenges contribute to what is recognized as one of the most significant constraints to the informed management of biodiversity, which is the lack of comprehensive knowledge on species distributions and their changes (Hortal *et al.*, 2015). This is particularly difficult to obtain when species identification is not straightforward (Chenuil *et al.*, 2019). In this context, the need for automatic species identification from media such as images has been highlighted by different authors (Gaston & O'Neill, 2004; MacLeod *et al.*, 2010).

Fuelled by recent advances in image classification techniques and the collection of large-scale image datasets, this new endeavour is finally taking shape, with various examples across distinct taxonomic groups in recent years (see review in: Wäldchen & Mäder, 2018). These studies typically use image-classification techniques based on the so-called deep-learning methods, which are capable of capturing high-level abstractions from data, and particularly on convolutional neural networks (CNNs), which are currently the state-of-the-art algorithm for image classification tasks. This type of artificial neural networks is particularly useful for processing data with a grid-like topology, such as images, and surpasses the need for preliminary feature extraction (that is, the selection of characteristics potentially important for classification), a common practice in the field prior to the development of these methods (Wäldchen & Mäder, 2018).

The application of deep-learning classification methods to biological images is a thriving field. A type of dataset in which these methods have been a turning point is that generated by camera traps (Chen *et al.*, 2014; Nguyen *et al.*, 2017; Norouzzadeh *et al.*, 2018; Miao *et al.*, 2019), but large image collections are becoming available in a variety of different contexts. Regarding taxonomy-related applications, a diverse array of studies focus on the identification of plants (mostly comprising vast taxonomic scopes) (Lee *et al.*, 2015; Zhou *et al.*, 2016; Barré *et al.*, 2017; Gogul & Kumar, 2017; Seeland *et al.*, 2019) or insects (Marques *et al.*, 2018; Arzar *et al.*, 2019; Almryad & Kutucu, 2020; Buschbacher *et al.*, 2020; Hansen *et al.*, 2020; Milošević *et al.*, 2020; Goodwin *et al.*, 2021), two taxonomic groups with a long tradition in automated identification. A few studies address aquatic wildlife such as fish (dos Santos & Goncalves, 2019; Rauf *et al.*, 2019; Lu *et al.*, 2020), corals (Gómez-Rios *et al.*, 2019a, b) and foraminifers (Hsiang *et al.*, 2019), while other taxonomic groups have been clearly under-represented. Apart from such case-specific applications, large-scale identification tools have been developed (Barré *et al.*, 2017; Buschbacher *et al.*, 2020), including also freely available mobile applications for the general public [e.g. iNaturalist Seek (https://www.inaturalist.org/pages/seek_app), Pl@ntNet (Affouard *et al.*, 2017) or Flora Incognita (Mäder *et al.*, 2021)]. These tools are revolutionizing the way that biodiversity is monitored and protected (Bonnet *et al.*, 2020).

Importantly, to our knowledge, all of these applications focus on taxa that are clearly morphologically distinct. The utility of deep-learning methods in classifying images of species with subtle morphological differences, which human observers have difficulties in identifying, remains to be investigated. Here we fill this gap by applying deep-learning tools to the classification of images belonging to a group of highly similar species: *Podarcis* Wagler, 1830 wall lizards. This genus of lacertids is among the most abundant and successful reptiles in the Mediterranean region. We focus on the particularly cryptic Iberian and North African group of species, also known as the *P. hispanicus* (Steindachner, 1870) species complex, which form a monophyletic clade within the genus (Salvi *et al.*, 2021; Yang *et al.*, 2021). Despite significant genetic and ecological differentiation (Caeiro-Dias *et al.*, 2018, 2021a, b; Pinho *et al.*, 2007, 2008), these species are notably hard to identify morphologically, as the high intraspecific diversity often obscures interspecific differences (Kaliontzopoulou *et al.*, 2012b). Identification of the correct species, therefore, requires expert intervention or genetic tools, and particular cases such as *P. lusitanicus* Geniez *et al.*, 2014 and *P. guadarramae* (Boscá, 1916) (until recently considered to be the same species) are cryptic (Caeiro-Dias *et al.*, 2021b; Geniez *et al.*, 2014). Wall lizards are currently models for biological studies in various disciplines (see: Salvi *et al.*, 2021). One species in this group (*P. carbonelli* Pérez Mellado, 1981) holds special conservation interest since it is classified as endangered and exhibits a declining population trend (Sá-Sousa *et al.*, 2009). Enabling straightforward identification of wall lizards is thus important for downstream studies involving their monitoring and conservation. At the same time, given their overall similarity, these species are also an adequate case study to address the advantages and limitations of deep-learning methods applied to identifying images of morphologically similar species.

MATERIAL AND METHODS

Classes considered

We focused on two distinct identification experiments. The first involved the discrimination between *P. bocagei* (Seoane, 1885) and *P. lusitanicus.* This species pair was chosen for this initial analysis because these two species occur together, often on the same walls, throughout the north-west of the Iberian Peninsula. Hence, distinguishing between them is usually of practical interest since the geographic location of collection or sighting cannot help in this case (as it often does with other species of wall lizards). Moreover, although generally similar, as any other species in the genus, these two taxa in particular show important differences in size, coloration and head shape (Kaliontzopoulou *et al.*, 2012a; Gomes *et al.*, 2016), which make them good candidates for a preliminary test of the usefulness of computer vision models for studying this system.

The second experiment involved separating nine groups, which encompassed diversity representative of the whole species complex. Therefore, this dataset includes a larger spectrum of morphological variation, including an array that goes from the more easily recognizable pair mentioned above, to cryptic forms, such as the pair of *P. lusitanicus* and *P. guadarramae* (Geniez *et al.*, 2014; Caeiro-Dias *et al.*, 2021b). The nine species/classes of *Podarcis* evaluated were: *P. bocagei*, *P. carbonelli*, *P. guadarramae*, *P. hispanicus*, *P. liolepis* (Boulenger, 1905), *P. lusitanicus*, *P. vaucheri* (Boulenger, 1905) *s.s.*, *P. vaucheri s.l.* and *P. virescens* Geniez *et al.*, 2014. *Podarcis vaucheri s.s.* includes the lineages 'PVMA' (from Morocco and Algeria), 'PVSSp' and 'PVSCSp' (from south and south-central Spain, respectively), whereas *P. vaucheri s.l.* includes the more distantly related lineages 'PHTA', 'PHBAT', 'PHAZA' and 'PHJS' from Tunisia, isolates in Algeria and southern Morocco as defined in (Kaliontzopoulou *et al.*, 2011). Recently, Bassitta *et al.* (2020) described a new species, *P. galerai* Basitta *et al.*, 2020, corresponding to the southern populations of *P. hispanicus*. However, a more recent evaluation (Yang *et al.*, 2021) demonstrated that *P. galerai* and *P. hispanicus* are sister-taxa, which contradicts one of the strongest arguments used by Bassitta *et al.* (2020) to separate the two forms. Hence, and also because splitting these two forms would imply a much lower sample size for both classes, in this study we preferred to adopt the conservative approach of merging the two putative species under the name *P. hispanicus*.

Image datasets

The images that were used as input for this work were obtained between 2005 and 2015 in the scope of studies related to the eco-evolutionary dynamics of Iberian and North African *Podarcis*. Some of these images have been the object of previous studies involving geometric morphometrics techniques or scale count data (e.g. Kaliontzopoulou *et al.*, 2012a, b). All individuals have been identified as one of the nine species/classes mentioned above via mtDNA sequencing at the population level (Kaliontzopoulou *et al.*, 2011; Caeiro-Dias *et al.*, 2018). Within each class, images were obtained in different collection localities (ranging from four to 21 localities per class, with more localities for species with a larger geographic distribution). The locality of origin was ignored in this work, but it adds to the natural variability of the data examined.

The images used are as standardized as possible (i.e. the same general perspectives were taken for all individuals against a low-complexity background) but they include substantial format, zoom, illumination and exposure differences, as well as positional variation and deformation, given that the animals were alive and moving when the photographs were taken. We used images from two perspectives: a dorsal view, including the entire body of the lizards (the tail, often autotomized, may or may not appear in the images) and a lateral close-up of the head. Figure 1 represents examples of these two perspectives (before pre-processing) for the same individual. Because of the marked sexual dimorphism present in this genus (Kaliontzopoulou *et al.*, 2007, 2015), males and females were analysed separately. Therefore, we analysed four distinct datasets for each of the two classification experiments (two different perspectives for each of the two sexes). Table 1 provides the breakdown of the total sample according to species, sex and perspective. The total number of images analysed ranged between 932 and 1101, depending on the dataset. Please note that the images used for discriminating *P. bocagei* and *P. lusitanicus* (244–313 images) were also part of this larger dataset.

Image pre-processing

We first made sure that images were all in the same orientation (snout facing towards the right), which required rotating a tiny percentage of images (most of the original dataset already conformed to this orientation). Subsequently, images were centred, cropped, converted to square format and resized to the same dimensions using the ImageMagick v.7.0.10 software (The ImageMagick Development Team, 2021). Although background removal is not mandatory for this type of analysis, many images included hand-written labels in the background (e.g. Fig. 1), potentially influencing classification outcomes. Instead of manually manipulating individual images to remove such labels, we removed the background from all images. This was performed automatically using ADOBE PHOTOSHOP 2021 (https://www.adobe.com/pt/products/photoshop.html) in batch mode, with some manual corrections when needed.

EXPERIMENTAL ANALYSES

We used the same general framework in the two classification experiments presented in this work. Prior to the analyses, the datasets were subdivided into five replicates of the same size and class frequencies for cross-validation, that is, we created datasets for five-fold cross-validation: three-folds (60% of images) were used for training, one-fold (20%) for validation and model parameter tuning during the learning process, and the remaining 20% were left unseen by the model for testing after the learning stage was completed.

We used the deep-learning library KERAS (Chollet *et al.*, 2018) with TensorFlow (Abadi *et al.*, 2016) as backend in PYTHON v.3.8. This is the most common framework for deep-learning image classification problems, enabling simple and streamlined workflows. Images were loaded with size 224 × 224 pixels. Besides rescaling the data so that all values fell between 0 and 1 (which is a common procedure also applied to validation and testing datasets), in the training datasets we performed data augmentation, since initial experiments suggested it greatly reduced overfitting. This is a technique that involves producing random modifications (such as rotation, zoom, range shifts, brightness changes, etc.) to the images presented to the model in order to create additional diversity. This procedure was different in the two classification experiments performed. In the two-species experiment, data-augmentation parameters (i.e. the top limits of the uniform distribution from which image modification values are drawn) were set as follows: rotation_range = 70, width_shift_range = height_shift_range = shear_range = zoom_range = 0.2, brightness_range between 0.5 and 1.5. In the nine-class experiment, because initial trials suggested overfitting was still a problem (training accuracy was always much higher than validation accuracy), we increased the diversity in the data presented to the model by increasing the range of some data-augmentation parameters, namely width_shift_range, height_shift_range, shear_range and zoom_range, which were set to 0.7, and brightness_range, which was established between 0.2 and 1.8. We did not augment data by flipping images since our datasets did not vary in this respect.

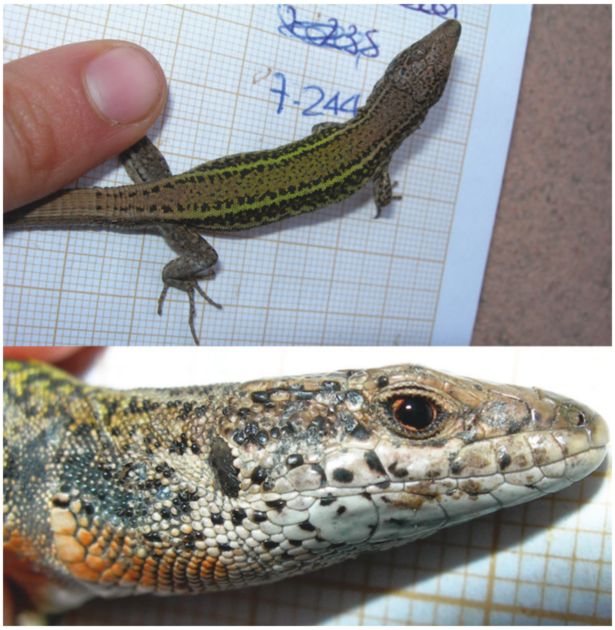


Figure 1. The two image types analysed in this study (before pre-processing): above, a dorsal view; below, a head lateral image. Both images correspond to the same *Podarcis vaucheri s.l.* male.

Table 1. Number of images per species, sex and image perspective

View	Females			Males		
	Dorsal	Head_lateral	Both	Dorsal	Head_lateral	Both
P. bocagei	168	171	167	210	214	210
P. carbonelli	96	95	95	108	108	108
P. guadarramae	49	49	49	63	63	63
P. hispanicus	31	31	31	41	41	41
P. liolepis	71	65	65	73	69	69
P. lusitanicus	76	76	76	98	99	98
P. vaucheri s.l.	49	51	49	61	61	61
P. vaucheri s.s.	206	233	202	216	241	215
P. virescens	186	186	185	205	205	205
TOTAL	932	957	919	1075	1101	1070

Concerning the deep-learning models used, we chose three architectures based on common approaches in the literature: InceptionV3 (Szegedy *et al.*, 2015), ResNet50 (He *et al.*, 2016) and Inception-ResNet-V2 (Szegedy *et al.*, 2017). InceptionV3 is based on the repetition of the 'Inception' module, which applies a 1 × 1 convolution and the concatenation of simultaneous operations to reduce the dimensionality of the problem, allowing an increase in the overall network depth. ResNet represented a dramatic breakthrough in the field by introducing the concept of residual learning, which involves adding identity mapping between some layers to improve backpropagation and minimize vanishing gradient problems, while also borrowing some concepts from Inception. Finally, Inception-ResNet-V2 combines the Inception architecture with residual learning, showing significant improvements both in training speed and in classification success. All of the networks chosen are readily available in KERAS. Using a common practice in the field, we initialized the models with weights pre-trained from ImageNet (a large image dataset developed for academic purposes; Deng *et al.*, 2009). The top fully connected layers were not imported. Instead, we added to the base model an average pooling layer, followed by a fully connected layer with 1024 units and ReLu activation, a 0.5 dropout step and a final classifier (a single unit using the sigmoid activation in the binary classification and nine units using softmax activation in the nine class problem). In all cases, the tuning of model parameters was carried out on the complete network (the original architectures and the top layers added) and was carried out using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001 (although in preliminary tests we performed with different learning rates). A cross-entropy loss function was used. The learning process was conducted for 1000 epochs with a batch size of 32 in the two-species classification experiment and of 2000 epochs and a batch size of 64 in the nine-class experiment. Because the classes in our datasets were unbalanced, and as preliminary runs showed an advantage in this approach, class weights were used to ensure that misclassifications in the lower frequency classes received higher penalties during model fitting. This was ensured by using the 'balanced' heuristic in the compute_class_weights function of the scikit-learn PYTHON module and using the resulting weights during model training. The classification success of the validation dataset for each cross-validation replicate was monitored during the learning stage, and used to guide the learning process and to select the best model. After training, learning curves for both the training and validation datasets were inspected using the TensorFlow visualization toolkit 'TensorBoard'.

The best models (that is, those with the best performance with respect to the validation set) obtained during the training stage were used to make predictions and evaluate the performance of the methods on each test set. A matrix of probabilistic predictions for the whole dataset (combining predictions for all five cross-validation replicates) was then used to evaluate each model individually and to produce model ensemble predictions. For this purpose, we combined the predictions of the three models for each image by calculating the arithmetic mean of the probability for each class to obtain a within-dataset model ensemble. We extended this approach by obtaining, for all individuals for which our dataset included both dorsal and head lateral images, the arithmetic mean of the probability estimated by all six models to produce a more representative prediction based on different image perspectives.

We used the PYTHON module scikit-learn to calculate several performance metrics: accuracy, precision, recall and F1-score, both globally using macro-averages and an estimate per class considering each class success versus all the others; and, in the case of the two-species experiment, the area under the receiver operating characteristic (ROC) curve, AUC. The confusion matrix detailing classification outcomes was also obtained. When relevant, performance metrics were compared using non-parametric procedures (Mann–Whitney–Wilcoxon tests in the case of independent samples, i.e. between datasets, and Wilcoxon signed-rank tests in the case of comparisons involving the same cross validation replicates, that is, within datasets). All scripts used for training and evaluation can be found in https://github.com/catpinho/image_classification.

Finally, we used Gradient-Weighted Class Activation Mapping, Grad-CAM (Selvaraju *et al.*, 2017) to produce heatmaps showing the areas of each training image that are important in classification. Grad-CAM is an increasingly popular visual explainer of deep-learning algorithms, particularly in the case of biological images, which uses the gradients of a class in a classification network flowing into the final convolution layer to produce a heatmap showing the visual localization of the important regions in the image involved in the classification. This was conducted for the best-performing model in each case. We followed the implementation suggested in https://keras.io/examples/vision/grad_cam/, with some minor modifications.

RESULTS

Discrimination between *Podarcis bocagei* and *P. lusitanicus*

The overall performance of the three methods for image classification of *P. bocagei* and *P. lusitanicus* in the four different datasets is shown in Table 2. Detailed results, including training, validation and test-set evaluation for all cross-validation sets, are shown in the Supporting Information, Tables S1–S4. Accuracy is generally high, ranging from 87.3% in the case of InceptionV3 in female dorsal images to 94.8% in male dorsal images when applying InceptionResNetV2. AUC ranges from 0.931 using InceptionV3 in female dorsal images to 0.984 using Inception-ResNetV2 in male dorsal and head lateral images. F1-scores show that, typically, *P. lusitanicus* is more frequently misclassified than *P. bocagei*, for both types of images and for both sexes. All three methods perform similarly in all datasets considering the three performance metrics. Identification of males is generally more accurate than that of females. Considering all five cross-validation replicates of the three models, the identification accuracy of males is significantly higher than that of females only when considering dorsal images ($P = 0.048$, Mann–Whitney–Wilcoxon test). The same result is obtained, but even more pronounced, using other metrics ($P = 0.009$ and $P = 0.030$ for AUC and F1-scores, respectively). With respect to head lateral images, the difference in identification accuracy between sexes also exists but it is significant only for differences in AUC ($P = 0.046$, Mann–Whitney–Wilcoxon test). There is no difference in performance using different image perspectives, neither in the case of males nor in that of females.

As an extension to this basic approach, we tested whether model ensembles (calculated by averaging predictions of different models) would increase classification success. Model ensembles within each of the four datasets do not always improve classification success compared to the best single model (see results in Table 2). For instance, in the case of head lateral images, prediction performance is worse with the model ensemble than when using the best-performing model alone. In the case of dorsal images, the improvement is slight for males and more substantial for females.

By contrast, combining the predictions from different views results in a much higher classification success in all cases, where accuracy reaches as high as 97.1% for males and 95.9% for females. These results are presented in Table 3 and the corresponding confusion matrices in Figure 2.

Grad-CAM heatmaps were produced only for the model showing the highest accuracy in each case (Inception-ResNet V2 in the case of male dorsal and head lateral images, ResNet50 in the case of female dorsal images and Inception V3 in the case of female head lateral images). Visualization of the heatmaps confirms that the models are indeed considering the lizard images for classification and not external features (like human fingers, writings, shadows and other non-lizard elements that appear in some images).

Examples of heatmaps used to discriminate the two classes are shown in Figure 3. In dorsal images, the model often uses the middle area of the trunk to discriminate the two classes. Still, the head region is also used (and both regions combined). In female dorsal images, the head is not as frequently used as the trunk, but the portion of the trunk used for discrimination is

Table 2. Evaluation of the three tested architectures in the four datasets for the two-class experiment. Models with the highest accuracy are highlighted in bold

Sex	View	Metric*	Inception V3	ResNet 50	Inception ResNetV2	Combined predictions
Males	Dorsal	Accuracy	0.935	0.922	**0.948**	0.955
		AUC	0.976	0.982	0.984	0.982
		F1 *Pboc*	0.951	0.941	0.962	0.967
		F1 *Plus*	0.905	0.887	0.919	0.930
	Head lateral	Accuracy	0.926	0.929	**0.936**	0.930
		AUC	0.972	0.975	0.984	0.976
		F1 *Pboc*	0.946	0.947	0.953	0.949
		F1 *Plus*	0.882	0.895	0.889	0.885
Females	Dorsal	Accuracy	0.873	**0.906**	0.905	0.943
		AUC	0.931	0.965	0.970	0.970
		F1 *Pboc*	0.905	0.930	0.929	0.948
		F1 *Plus*	0.810	0.857	0.859	0.908
	Head lateral	Accuracy	**0.935**	0.919	0.927	0.911
		AUC	0.962	0.959	0.976	0.972
		F1 *Pboc*	0.953	0.941	0.947	0.937
		F1 *Plus*	0.897	0.87	0.872	0.847

*AUC refers to the area under the ROC curve. *Podarcis bocagei* was the positive class.
F1, harmonic mean of precision and recall; *Pboc*, *Podarcis bocagei*; *Plus*, *Podarcis lusitanicus*.

Table 3. Classification success of the six-model ensembles for the two-class experiments

	Males					Females				
	Accuracy	AUC	F1	Precision	Recall	Accuracy	AUC	F1	Precision	Recall
Average/global	0.971	0.997	0.966	0.970	0.962	0.959	0.992	0.952	0.952	0.952
P. bocagei			0.979	0.972	0.986			0.970	0.970	0.970
P. lusitanicus			0.953	0.968	0.939			0.934	0.934	0.934

AUC was calculated assuming *P. bocagei* as the positive case; F1, precision and recall were macro-averaged.

generally more anterior than in males. In both male and female head lateral images, the area around the ear is the one most frequently used for classification, although this region could be more or less shifted towards the throat in both sexes.

Discrimination between the nine groups

Overall, the performance of the different models for classification of the nine classes is worse than in the two-class case. Unlike the experiment involving only *P. bocagei* and *P. lusitanicus*, in all analyses considering nine classes there is some evidence of overfitting (see the Supporting Information, Tables S5–S9 for detailed training, validation and testing evaluation scores), which could not be completely overcome by varying the hyperparameters. A summary of the performance of each model is presented in Table 4.

In general, accuracy ranges from 76.3% for ResNet50 in female head perspectives to 85.3% for InceptionResNetV2 in male dorsal views. A striking result is the highly significant difference between male and female image identification accuracy, with consistently higher accuracies in male datasets, which holds for both types of images ($P < 0.0001$ for all comparisons, both for accuracy and F1 score, Mann–Whitney–Wilcoxon test). On the other hand, there are no differences in performance between the two types of images, neither for males nor for females. There are also no major differences between models in classification success. The only significant difference is detected in female head lateral images, in which ResNet50 performs significantly worse than Inception-ResNet V2 ($P = 0.0325$ for both accuracy and F1-score, Wilcoxon signed rank test). Unlike the two-class case, in which the utility of ensemble models is mostly restricted to the combination of predictions from different perspectives, without important improvements in the within-dataset case, in the nine-class experiment ensemble models combining predictions from the three architectures for each image perspective greatly improve classification accuracy when compared to the best single model (see Table 4).

Using estimates from different views by averaging across the six model predictions improves classification success even further. These results are shown in Table 5 and the respective confusion matrices shown in Figure 4. In this case, prediction accuracy reaches as high as 93.5% for males and 91.2% for females.

The distribution of classification metrics according to the species is shown in Table 5. Taking a deeper look into these classification scores, it appears that several species are fairly well recognizable, with F1 scores above 0.90: this is the case for *P. bocagei*, *P. carbonelli*, *P. lusitanicus*, *P. vaucheri s.l.*, *P. vaucheri s.s.* and *P. virescens* in males, and for the same species except *P. lusitanicus* in females. The most problematic species is, in both sexes, *P. liolepis*. Considering confusion matrices, it is noticeable that individuals of this species are often misclassified as *P. virescens* (more so in the case of females than males). Noteworthy is that the misclassification between the cryptic *P. guadarramae* and *P. lusitanicus* is minimal (7.9% of *P. lusitanicus* females and 4.6% of males are classified as *P. guadarramae* and 0 and 1.6% of *P. guadarramae* females and males are classified as *P. lusitanicus*; see Fig. 4).

As for the two-class problem, Grad-CAM analyses show that, typically, the models use lizard – and not other – features for classification. However, even with the visualization tool available, it is not straightforward to understand what the model considers for discrimination. More precisely, the same regions seem to be used to classify distinct species, and it is not evident how differences in these regions are used. The most common patterns for each species are summarized in Tables 6 and 7 (for males and females, respectively).

DISCUSSION

The possibility of automating the identification of biological images can bring exciting new perspectives for the study and monitoring of biodiversity by surpassing the need of expert intervention and reducing the expenses associated to alternative techniques, such as molecular tools, but also by potentially signalling morphological differences that may have remained

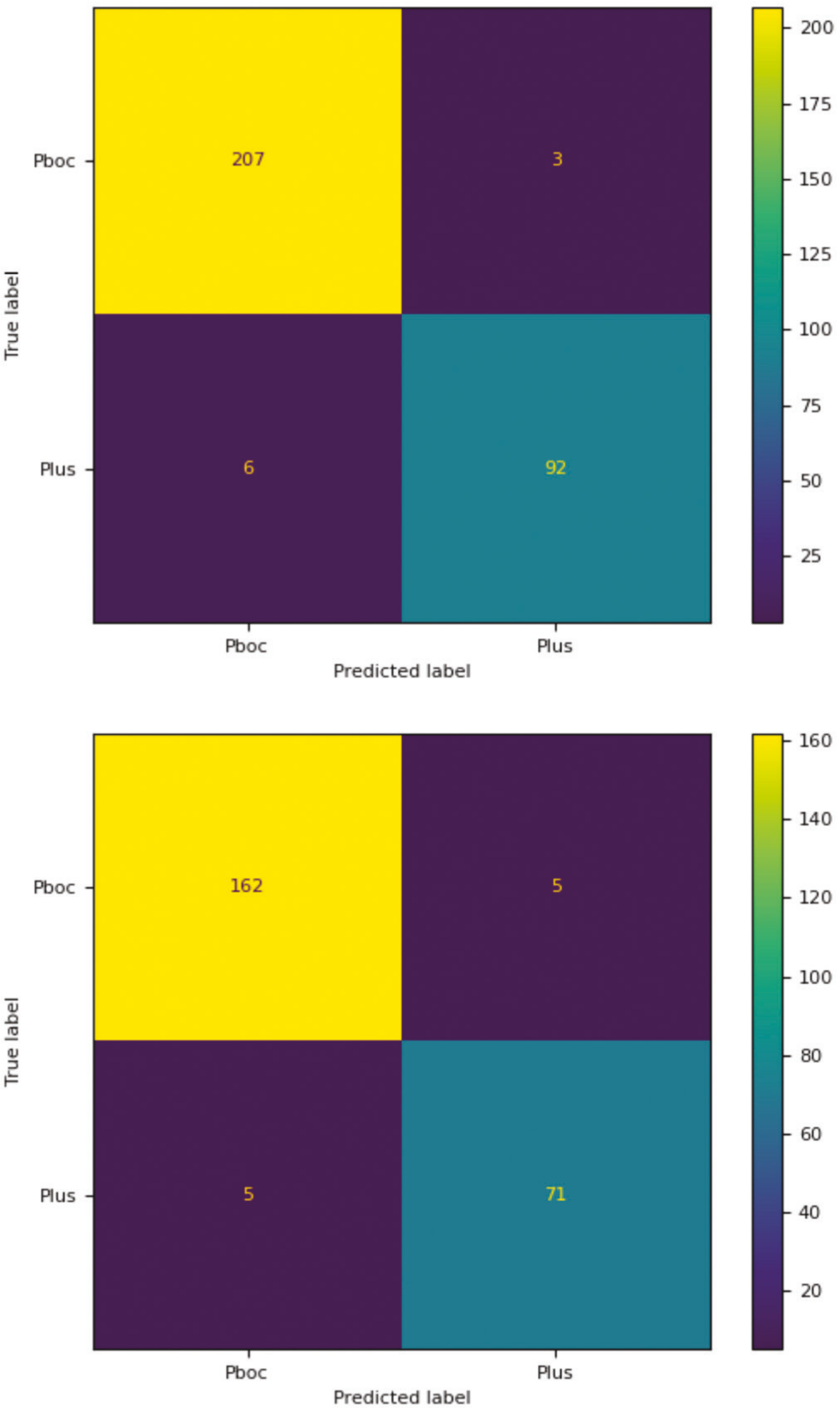


Figure 2. Confusion matrix for male (upper) and female (lower) image classification for the two-class case based on a combination of predictions from six models. Abbreviations used: Pboc, *P. bocagei*; Plus, *P. lusitanicus*.

elusive to human observers. In this study, we proposed the application of deep-learning algorithms to identify images belonging to closely related and morphologically similar lizard species. To our knowledge, this is one of the first studies with a taxonomic endeavour conducted in squamates, and the first not focused on snakes. The objects of this study, wall lizards belonging to the *P. hispanicus* species complex, are common, widespread

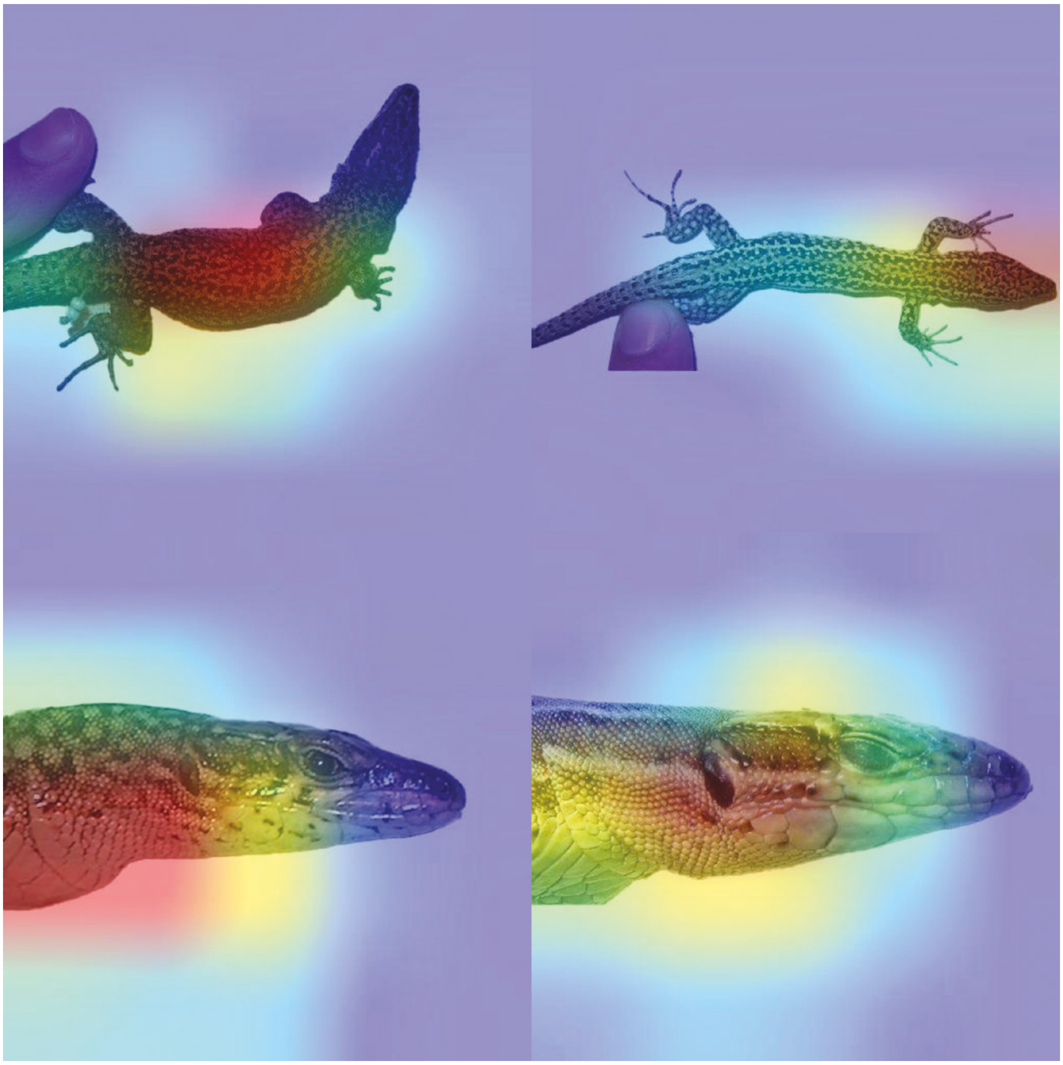


Figure 3. Example of Grad-CAM heatmaps obtained for *Podarcis lusitanicus*. The upper images show two common patterns observed in male dorsal images (also found, albeit with some differences, in females). The bottom images exhibit the patterns most frequently found in male and female head lateral images (here illustrated in two females).

Table 4. Evaluation of the three tested architectures and their combined predictions in the four datasets for the nine-class experiment. Models with the highest accuracy are highlighted in bold

Sex	View	Metric	Inception V3	ResNet 50	Inception ResNetV2	Combined predictions
Males	Dorsal	Accuracy	0.849	0.832	**0.853**	0.886
		F1 macro	0.826	0.806	0.829	0.866
	Head lateral	Accuracy	0.832	**0.840**	0.828	0.876
		F1 macro	0.811	0.817	0.807	0.854
Females	Dorsal	Accuracy	0.799	0.783	**0.819**	0.854
		F1 macro	0.772	0.758	0.793	0.827
	Head lateral	Accuracy	0.783	0.763	**0.804**	0.830
		F1 macro	0.744	0.727	0.775	0.802

Table 5. Classification success of the six-model ensembles for the nine-class experiment

	Males				Females			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
Average/global	0.935	0.923	0.940	0.910	0.912	0.894	0.918	0.877
P. bocagei		0.965	0.958	0.971		0.944	0.925	0.964
P. carbonelli		0.986	1.000	0.972		0.943	0.929	0.958
P. guadarramae		0.884	0.864	0.905		0.866	0.875	0.857
P. hispanicus		0.883	0.944	0.829		0.852	0.867	0.839
P. liolepis		0.829	0.944	0.739		0.793	0.957	0.677
P. lusitanicus		0.911	0.935	0.888		0.863	0.952	0.789
P. vaucheri s.l.		0.983	1.000	0.967		0.936	0.978	0.898
P. vaucheri s.s.		0.945	0.924	0.967		0.934	0.923	0.946
P. virescens		0.920	0.890	0.951		0.911	0.861	0.968

F1, harmonic mean of precision and recall.

species frequently found across Iberia and the Maghreb, and have long been a challenge for taxonomists and naturalists alike, because of the combination of low interspecific with huge inter-individual morphological variation (Kaliontzopoulou *et al.*, 2012b). Our results highlight how computer vision models can be an important addition to the taxonomist toolbox even in the case of species that are difficult to identify. However, they also bring into focus relevant challenges associated with the biological interpretation of the information that such models provide; in particular, with respect to the morphological features considered to differentiate the focal species.

Telling cryptic species apart: overall high classification success

Throughout, and particularly for researchers familiar with the difficulties of identifying cryptic species, the results obtained here are remarkable as puzzling. With respect to the first classification experiment we conducted, the performance of computer vision models was remarkably high. This was expected, since the two species investigated, *P. bocagei* and *P. lusitanicus*, show morphological differences that enable their distinction by experts (namely a smaller size, less intense green coloration in the dorsum and flatter heads in the case of *P. lusitanicus*), but compared to other species that have been the object of studies involving deep-learning tools, they can still be considered fairly similar. In this context, the high classification success obtained in this study (from 90.4% in female dorsal to 94.8% accuracy in male dorsal images for single models, and as high as 97.1% and 95.9% for males and females, respectively, using ensemble models to combine the results from the two image perspectives) is comparable to the accuracies generally reported in similar studies involving species not considered cryptic.

When addressing the more complex problem of simultaneously distinguishing among the nine species in the *P. hispanicus* species complex, classification success dropped significantly. For single models, accuracy ranged from 80.4% to 85.3% for the best-performing architectures applied to female head lateral and male dorsal images, respectively. Combining the predictions obtained with different architectures for each image perspective increased this success to a moderate extent, and classification was highest when combining predictions from all six different image perspectives and model architectures for the same individual (93.5% in the case of males and 91.2% in the case of females). Although these improvements were significant, these accuracies are still below those obtained for the simpler distinction between *P. bocagei* and *P. lusitanicus*. However, it should be emphasized that, even if the models appear to fail often at classifying individuals into species among the nine classes compared to the two-class case or compared to computer vision models applied to other systems, the results obtained in this study are still, by far, the highest classification success obtained applying morphological characters in this system. Kaliontzopoulou *et al.* (2012b) focused on the same species group and applied a classification scheme based on classical characters traditionally used to distinguish lacertids and other lizard species: biometry (linear body measurements), pholidotic (scale-count) characters (in some cases obtained by analysing the same images used in the present study) and a combination of both types of characters. Although the exact classes considered were not the same and hence the results cannot be directly compared, classification results were typically much worse (mean of 56.6% in males and 51.73% in females). Results improved when other classification schemes were considered (binary schemes involving one class vs. all the others or all pairwise comparisons), which

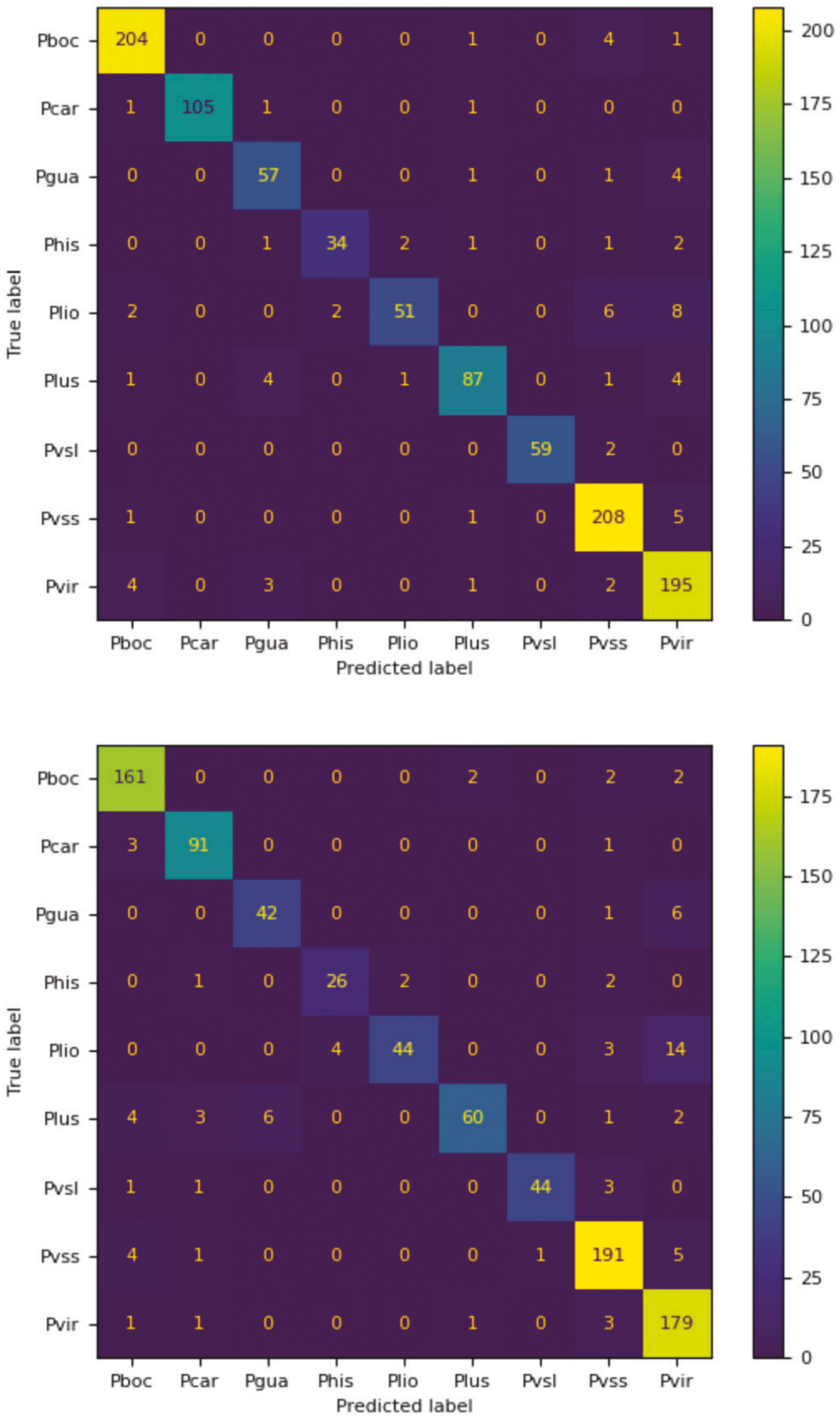


Figure 4. Confusion matrix for male (upper) and female (lower) image classification for the nine-class experiment based on a combination of predictions from six models. Abbreviations used: Pboc, *P. bocagei*; Pcar, *P. carbonelli*; Phis, *P. hispanicus*; Plio, *P. liolepis*; Plus, *P. lusitanicus*; Pvsl, *P. vaucheri s.l.*; Pvss, *P. vaucheri s.s.*; Pvir, *P. virescens*.

Table 6. Summary of Grad-CAM results for each class (males)

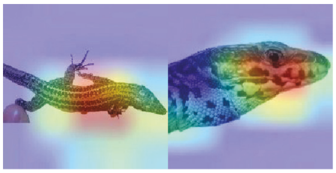
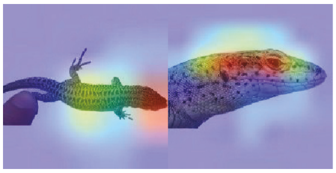
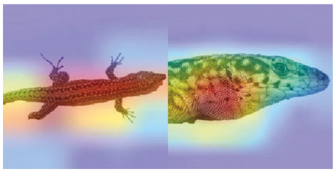
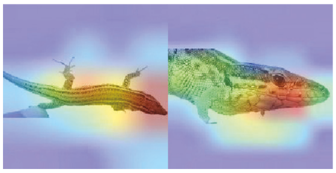
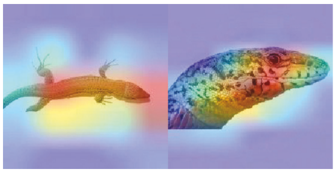
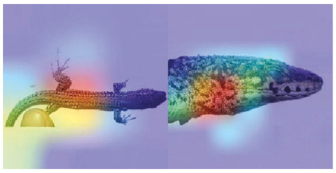
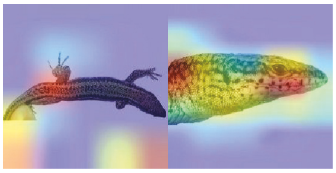
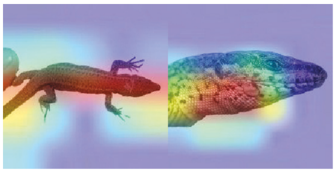
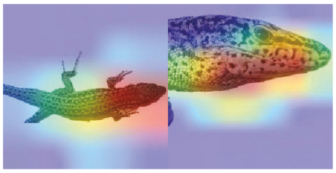
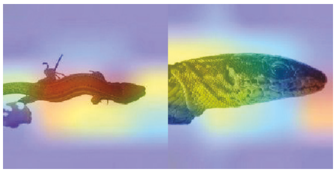
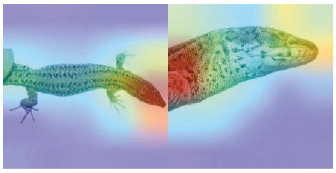
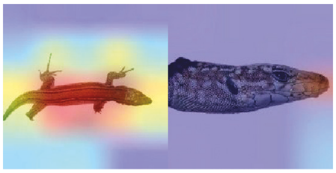
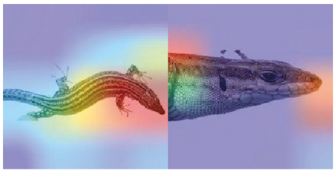
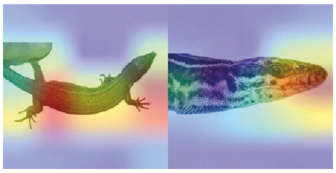
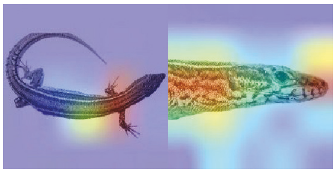
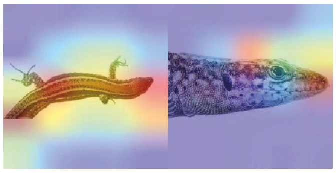
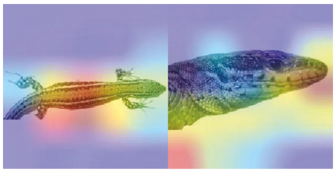
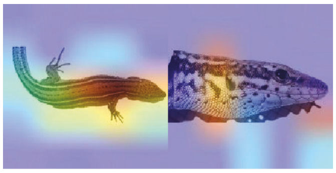
P. bocagei	Highly variable (no clear pattern). All portions of the dorsal views were equally used. In head images the area around the eye, the top of the head, the snout and the throat were all used in similar proportions.	
P. carbonelli	Variable for both views. Snout and middle of the dorsum used in dorsal view. Top of the head most frequently (but not strictly) used in lateral view.	
P. guadarramae	Whole body used for dorsal view (but variable); either throat (most common) or ear region used in head lateral views.	
P. hispanicus	Variable. Anterior portion of snout used more frequently than in other species for both dorsal and head lateral views.	
P. liolepis	Highly variable. Whole body used in most dorsal images, area around the eye and throat used in head lateral views, but other patterns common.	
P. lusitanicus	Highly variable. All parts of the dorsum used (but frequently the most posterior part); area around the ear frequently used in head lateral images.	
P. tunesiacus	Highly variable. Dorsal area near the insertion of the posterior limbs used more frequently than in other species; different regions of the head used, often simultaneously.	
P. vaucheri	Highly variable. Different regions of dorsum (from head to the posterior region) used in dorsal images, all portions of the head, but most frequently the throat, used in lateral images.	
P. virescens	Highly variable. All parts of both images used. Head and anterior part of the dorsum more used than in other species.	

Table 7. Summary of Grad-CAM results for each class (females)

P. bocagei	Highly variable. Mid-portion of the dorsum used frequently (although other areas as well). Tip of the snout used often, but area around the ear and throat are also relevant.	
P. carbonelli	Variable. In the dorsal view, the tip of the snout is frequently used. In the head lateral view, the tip of the snout is also commonly used, as well as the most posterior region of the head.	
P. guadarramae	Variable. Mid portion of the dorsum and tip of the snout are the regions used more frequently in dorsal and head lateral views, respectively.	
P. hispanicus	Variable. The head and most anterior part of the dorsum are frequently used in the dorsal view. Snout and/or top of posterior region of head used.	
P. liolepis	Variable. Different parts of the dorsum are used, whereas the tip of the snout is used in most head lateral images.	
P. lusitanicus	Anterior dorsum, in the dorsal view, and both snout and posterior side of the head (in head lateral views) frequently used.	
P. tunesiacus	Variable. Tip of the snout and posterior part of the trunk more used than in other species; snout and top head region behind the eye used with some frequency.	
P. vaucheri	Highly variable. All parts of the dorsum used in dorsal images, various parts of the head (but frequently snout and throat combined) used in head lateral images.	
P. virescens	Highly variable. All portions of the dorsum used in dorsal images, region around and behind the ear more used than in other species for head lateral images.	

could also be a future possibility to consider improving even further the use of computer vision models.

Identification of particular species: successes and failures

The reduced performance of models in the nine-class problem is largely driven by the moderate to low ability of the models to classify certain species. In fact, whereas species such as *P. virescens*, *P. bocagei* or *P. carbonelli* are typically successfully classified (particularly, but not only, using model ensembles), individuals from other species are also frequently misidentified, and this is not completely solved by combining predictions from different model architectures and image perspectives. *Podarcis liolepis*, the Catalonian wall lizard, is a case in point. This species has a wide distribution throughout the eastern half of the Iberian Peninsula, probably one of the most widespread among those included in this study (Renoult *et al.*, 2010). This means that this species is prone to encompass high morphological variability resulting from mechanisms such as adaptation and developmental plasticity to cope with widely varying environmental conditions, as has been observed in other species of the same clade with more limited distribution ranges (Kaliontzopoulou *et al.*, 2010a, b, 2018). Moreover, some southern populations are completely isolated from the remainder of the species, suggesting that genetic drift might accentuate differentiation among populations. Finally, unlike other lizard species, which tend to be more or less homogeneous genetically, *P. liolepis* includes two distinct mitochondrial DNA lineages, one of them resulting from introgression with a now extinct form (Renoult *et al.*, 2009). It is possible that these complex evolutionary dynamics have left their mark on morphological variation, making *P. liolepis* more diverse, in some aspects, than other species of the complex, hence more difficult to classify. Another possibly important reason for the misclassification of *P. liolepis* and other species is also current gene flow. This is a feature common in all *Podarcis* (Yang *et al.*, 2021), and the *P. hispanicus* complex is no exception (see e.g. Caeiro-Dias *et al.*, 2021a). This phenomenon could have a strong impact on the morphology of some individuals, particularly those coming from regions near contact zones.

An unexpected result of this study is the relatively high ability (considering the prior expectations) for the models to distinguish between *P. lusitanicus* and *P. guadarramae*. This is the most cryptic species pair included in this study, as individuals of these two species cannot be told apart even by the most experienced experts (Geniez *et al.*, 2014). It is thus remarkable that in our study the proportion of individuals of one species identified as the other is substantially low (0.0–7.9% using model ensembles). Future studies may attempt to extract and evaluate the models' features for classification and assess their eco-evolutionary implications (see also below).

Another remarkable result is the typically high classification success obtained for *Podarcis carbonelli*. Amongst the species in the Iberian and North African group, this is the only one that is currently of conservation concern, having been classified as 'endangered' by the IUCN (Sá-Sousa *et al.*, 2008). It is thus a promising result that computer vision models can identify *P. carbonelli* with a low error rate, since it enables the possibility of establishing citizen science distribution monitoring programmes directed at this species once these models become available in naturalist mobile applications.

Sexual dimorphism and classification success

Lizards of the genus *Podarcis* typically exhibit a marked sexual dimorphism, which is also accompanied by a tendency for differences between different species to be more pronounced in males than in females. Indeed, females are usually more uniform since they are less brightly coloured and lack other external features that typically help identify males (Kaliontzopoulou *et al.*, 2007). Similar to the difficulties experienced by human observers, classification success in this study was lower in females than in males in both problems. In the two-species experiment, differences in classification success between the sexes were only evident in the dorsal view, whereas in the nine-species experiment these differences were significant for both views. The fact that sexual dimorphism does not affect the distinction between head lateral images of *P. bocagei* and *P. lusitanicus* probably results from the fact that the most obvious difference among the two species, the high degree of head flattening exhibited by *P. lusitanicus*, which is probably related to adaptation to living in rock crevices (Gomes *et al.*, 2016; Kaliontzopoulou *et al.*, 2012a, b), is shared by both males and females, and Grad-CAM analysis suggests that the height of the head is indeed a feature that the models consider. However, this feature alone does not distinguish between females of all nine classes, which is reflected in this case in a much lower ability of the models to correctly classify females in general.

Curiously, despite different levels of classification success between the sexes, overall patterns of classification and confusion between species appear to be concordant, where, for example, *P. liolepis* is poorly classified both in the case of males and females, and with similar percentages classified as other species. This suggests that at least some of the features that the models are using are not sexually dimorphic.

The link between morphology and classification: explaining deep-learning models

The visualization of heatmaps produced by Grad-CAM allowed highlighting regions in the images used by the algorithms to distinguish between species. On the one hand, this was important to verify that the models were using features of the morphology of lizards and not irrelevant regions of the images. However, the attempt to decipher the features in lizard images that differed the most between species (and that could constitute valuable new knowledge in terms of understanding the ecoevolutionary dynamics of these species) was hampered by the great diversity of patterns found within each species coupled with the repetition of the same regions in different species – highlighting that, probably, different aspects of these regions were used for classification but did not provide hints on which particular aspects these were. This could, in fact, be a fruitful area of future method development for collaborative research between biologists and analysts specialized in computer vision and machine-learning algorithms to aid the extraction of biologically relevant information from this kind of model. Although this analysis may lack some objectivity, since summarizing heatmap results is not straightforward, it appears that models use more diverse regions of the body to classify males than to classify females. This is in line with the trends described in the previous section, highlighting that male exhibit more differences between species than females.

Methodological considerations

Despite our best efforts, classification success for the nine-class experiment was still problematic in some cases. Although there are probable biological causes for this pattern (see above), we cannot rule out that methodological issues involving sampling or model implementation are behind this suboptimal result. A possibly relevant aspect involves the unbalanced sample sizes of the different classes. We tried to minimize the impact of this problem in our workflow, but our results suggest that the species with the worst classification success are also those with the lowest sample sizes (a well-known problem in deep learning; Liu *et al.*, 2017). Therefore, adding images of these species from other sources (like citizen science platforms) to increase sample sizes is an important future addition to this work. Similarly, since classification success improved when combining information from different perspectives, it is possible that adding photographs from different views (e.g. ventral scales, the gular area, etc.) improves even further the classification outcomes.

An interesting observation from this study is that there are no significant differences in success when applying different deep-learning models. The three models used in this work differ in depth and in the general architecture of the convolutional neural networks. Despite these differences, they all perform similarly on these datasets. However, even if the overall result is the same, it does not mean that the models are considering the same features of the images for classification. Combining the three models by performing a simple average of the predictions already improves classification success in the nine-class problem, but it involves a considerable computational cost and may not be feasible for general analyses in the long term.

Because of moderate classification accuracy for some species, the practical deployment of these models for research or conservation purposes is still not a possibility. However, the challenge of discriminating between all nine classes simultaneously is an interesting academic exercise but is not a realistic problem that a naturalist will face in the field; although there are species that overlap and regions where the distribution is not well-known, a real-life problem in this species complex will involve distinguishing between, at most, three to four species simultaneously. Including geographical information to assist classification and/or to reduce the number of classes under consideration will thus certainly facilitate the classification problem, and the results obtained here for the two-species classification experiment provide a particularly promising way forward.

CONCLUSIONS

With this work we illustrate that deep-learning models can be successfully used to identify wall lizard species, achieving classification accuracies probably comparable to those of experienced observers and much higher than those of the common citizen. Moreover, beyond the specific problem of classifying wall lizards, this work shows that computer vision models can be useful for the visual distinction of cryptic species, something that had remained unexplored in the literature, thus opening promising research and application avenues. This includes the case of species such as *P. lusitanicus* and *P. guadarramae*, for which this work is the first suggesting morphological differences.

ACKNOWLEDGEMENTS

The authors would like to thank G. Caeiro-Dias for providing photographs for this study and D. Cunha for assistance in removing image backgrounds. CP is supported by National Funds through FCT-Fundação para a Ciência e a Tecnologia in the scope of the project UIDB/50027/2020. AK is supported by a Ramón y

Cajal research grant co-funded by the Spanish State Research Agency and the European Social Fund (RYC2019-026688-I/AEI/10.13039/501100011033).

DATA AVAILABILITY

The image dataset used in this study can be accessed in the following link: https://tinyurl.com/Podarcis-images.

REFERENCES

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M. 2016. TensorFlow: a system for Large-Scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16), 265–283.

Affouard A, Goëau H, Bonnet P, Lombardo JC, Joly A. 2017. Pl@ntnet app in the era of deep learning. ICLR: International Conference on Learning Representations, Apr 2017, Toulon, France. ffhal-01629195f.

Almryad AS, Kutucu H. 2020. Automatic identification for field butterflies by convolutional neural networks. *Engineering Science and Technology, an International Journal* **23**: 189–195.

Arzar NNK, Sabri N, Johari NFM, Shari AA, Noordin MRM, Ibrahim S. 2019. Butterfly species identification using convolutional neural network (CNN). 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS). IEEE, 221–224.

Barré P, Stöver BC, Müller KF, Steinhage V. 2017. LeafNet: a computer vision system for automatic plant species identification. *Ecological Informatics* **40**: 50–56.

Bassitta M, Buades JM, Pérez-Cembranos A, Pérez-Mellado V, Terrasa B, Brown RP, Navarro P, Lluch J, Ortega J, Castro JA. 2020. Multilocus and morphological analysis of south-eastern Iberian wall lizards (Squamata, Podarcis). *Zoologica Scripta* **49**: 668–683.

Bonnet P, Joly A, Faton JM, Brown S, Kimiti D, Deneu B, Servajean M, Affouard A, Lombardo JC, Mary L. 2020. How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. *Ecological Solutions and Evidence* **1**: e12023.

Buschbacher K, Ahrens D, Espeland M, Steinhage V. 2020. Image-based species identification of wild bees using convolutional neural networks. *Ecological Informatics* **55**: 101017.

Caeiro-Dias G, Luís C, Pinho C, Crochet PA, Sillero N, Kaliontzopoulou A. 2018. Lack of congruence of genetic and niche divergence in *Podarcis hispanicus* complex. *Journal of Zoological Systematics and Evolutionary Research* **56**: 479–492.

Caeiro-Dias G, Brelsford A, Kaliontzopoulou A, Meneses-Ribeiro M, Crochet PA, Pinho C. 2021a. Variable levels of introgression between the endangered *Podarcis carbonelli* and highly divergent congeneric species. *Heredity* **126**: 463–476.

Caeiro-Dias G, Rocha S, Couto A, Pereira C, Brelsford A, Crochet PA, Pinho C. 2021b. Nuclear phylogenies and genomics of a contact zone establish the species rank of *Podarcis lusitanicus* (Squamata, Lacertidae). *Molecular Phylogenetics and Evolution* **164**: 107270.

Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM. 2015. Accelerated modern human–induced species losses: entering the sixth mass extinction. *Science Advances* **1**: e1400253.

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

Chenuil A, Cahill A, Délémontey N, du Luc E, Fanton H. 2019. Problems and questions posed by cryptic species. A framework to guide future studies. *From assessing to conserving biodiversity*. Cham: Springer, 77–106.

Chollet F. 2018. Keras: the python deep learning library. *Astrophysics Source Code Library*: ascl-1806. https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C/abstract.

Costello MJ, May RM, Stork NE. 2013. Can we name earth's species before they go extinct? *Science* **339**: 413–416.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. Imagenet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 248–255.

Engel MS, Ceríaco LMP, Daniel GM, Dellapé PM, Löbl I, Marinov M, Reis RE, Young MT, Dubois A, Agarwal I, Lehmann AP, Alvarado M, Alvarez N, Andreone F, Araujo-Vieira K, Ascher JS, Baêta D, Baldo D, Bandeira SA, Barden P, Barrasso DA, Bendifallah L, Bockmann FA, Böhme W, Borkent A, Brandão CRF, Busack SD, Bybee SM, Channing A, Chatzimanolis S, Christenhusz MJM, Crisci JV, D'Elía G, Da Costa LM, Davis SR, Lucena CASD, Deuve T, Fernandes Elizalde S, Faivovich J, Farooq H, Ferguson AW, Gippoliti S, Gonçalves FMP, Gonzalez VH, Greenbaum E, Hinojosa-Díaz IA, Ineich I, Jiang J, Kahono S, Kury AB, Lucinda PHF, Lynch JD, Malécot V, Marques MP, Marris JWM, Mckellar RC, Mendes LF, Nihei SS, Nishikawa K, Ohler A, Orrico VGD, Ota H, Paiva J, Parrinha D, Pauwels OSG, Pereyra MO, Pestana LB, Pinheiro PDP, Prendini L, Prokop J, Rasmussen C, Rödel M-O, Rodrigues MT, Rodríguez SM, Salatnaya H, Sampaio I, Sánchez-García A, Shebl MA, Santos BS, Solórzano-Kraemer MM, Sousa ACA, Stoev P, Teta P, Trape J-F, Dos Santos CV-D, Vasudevan K, Vink CJ, Vogel G, Wagner P, Wappler T, Ware JL, Wedmann S, Zacharie CK. 2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society* **193**: 381–387.

Gaston KJ, O'Neill MA. 2004. Automated species identification: why not? *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **359**: 655–667.

Geniez P, Sa-Sousa P, Guillaume CP, Cluchier A, Crochet PA. 2014. Systematics of the *Podarcis hispanicus* complex

(Sauria, Lacertidae) III: valid nomina of the western and central Iberian forms. *Zootaxa* **3794**: 1–51.

Gogul I, **Kumar VS. 2017**. Flower species recognition system using convolution neural networks and transfer learning. 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN). IEEE, 1–6.

Gomes V, **Carretero MA**, **Kaliontzopoulou A. 2016**. The relevance of morphology for habitat use and locomotion in two species of wall lizards. *Acta Oecologica* **70**: 87–95.

Gómez-Ríos A, **Tabik S**, **Luengo J**, **Shihavuddin ASM**, **Herrera F. 2019a**. Coral species identification with texture or structure images using a two-level classifier based on Convolutional Neural Networks. *Knowledge-Based Systems* **184**: 104891.

Gómez-Ríos A, **Tabik S**, **Luengo J**, **Shihavuddin ASM**, **Krawczyk B**, **Herrera F. 2019b**. Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation. *Expert Systems with Applications* **118**: 315–328.

Goodwin A, **Padmanabhan S**, **Hira S**, **Glancey M**, **Slinowsky M**, **Immidisetti R**, **Scavo L**, **Brey J**, **Sai Sudhakar BMM**, **Ford T**, **Heier C**, **Linton YM**, **Pecor DB**, **Caicedo-Quiroga L**, **Acharya S. 2021**. Mosquito species identification using convolutional neural networks with a multitiered ensemble model for novel species detection. *Scientific Reports* **11**: 13656.

Hansen OL, **Svenning JC**, **Olsen K**, **Dupont S**, **Garner BH**, **Iosifidis A**, **Price BW**, **Høye TT. 2020**. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and Evolution* **10**: 737–747.

He K, **Zhang X**, **Ren S**, **Sun J. 2016**. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Hopkins GW, **Freckleton RP. 2002**. Declines in the numbers of amateur and professional taxonomists: implications for conservation. *Animal Conservation* **5**: 245–249.

Hortal J, **de Bello F**, **Diniz-Filho JAF**, **Lewinsohn TM**, **Lobo JM**, **Ladle RJ. 2015**. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**: 523–549.

Hsiang AY, **Brombacher A**, **Rillo MC**, **Mleneck-Vautravers MJ**, **Conn S**, **Lordsmith S**, **Jentzen A**, **Henehan MJ**, **Metcalfe B**, **Fenton IS**, **Wade BS**, **Fox L**, **Meilland J**, **Davis CV**, **Baranowski U**, **Groeneveld J**, **Edgar KM**, **Movellan A**, **Aze T**, **Dowsett HJ**, **Miller CG**, **Rios N**, **Hull PM**. **2019**. Endless forams: > 34 000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. *Paleoceanography and Paleoclimatology* **34**: 1157–1177. https://doi.org/10.1029/2019PA003612.

Kaliontzopoulou A, **Carretero MA**, **Llorente GA. 2007**. Multivariate and geometric morphometrics in the analysis of sexual dimorphism variation in *Podarcis* lizards. *Journal of Morphology* **268**: 152–165.

Kaliontzopoulou A, **Carretero MA**, **Llorente GA. 2010a**. Intraspecific ecomorphological variation: linear and geometric morphometrics reveal habitat-related patterns within *Podarcis bocagei* wall lizards. *Journal of Evolutionary Biology* **23**: 1234–1244.

Kaliontzopoulou A, **Carretero MA**, **Sillero N. 2010b**. Geographic patterns of morphological variation in the lizard *Podarcis carbonelli*, a species with fragmented distribution. *Herpetological Journal* **20**: 41–50.

Kaliontzopoulou A, **Pinho C**, **Harris DJ**, **Carretero MA. 2011**. When cryptic diversity blurs the picture: a cautionary tale from Iberian and North African *Podarcis* wall lizards. *Biological Journal of the Linnean Society* **103**: 779–800.

Kaliontzopoulou A, **Adams DC**, **van der Meijden A**, **Perera A**, **Carretero MA. 2012a**. Relationships between head morphology, bite performance and ecology in two species of *Podarcis* wall lizards. *Evolutionary Ecology* **26**: 825–845.

Kaliontzopoulou A, **Carretero MA**, **Llorente GA. 2012b**. Morphology of the *Podarcis* wall lizards (Squamata: Lacertidae) from the Iberian Peninsula and North Africa: patterns of variation in a putative cryptic species complex. *Zoological Journal of the Linnean Society* **164**: 173–193.

Kaliontzopoulou A, **Carretero MA**, **Adams DC. 2015**. Ecomorphological variation in male and female wall lizards and the macroevolution of sexual dimorphism in relation to habitat use. *Journal of Evolutionary Biology* **28**: 80–94.

Kaliontzopoulou A, **Pinho C**, **Martínez-Freiría F. 2018**. Where does diversity come from? Linking geographical patterns of morphological, genetic, and environmental variation in wall lizards. *BMC Evolutionary Biology* **18**: 124.

Kingma DP, **Ba J. 2014**. Adam: a method for stochastic optimization. *arXiv preprint arXiv*: 1412.6980.

Lee SH, **Chan CS**, **Wilkin P**, **Remagnino P. 2015**. Deep-plant: plant identification with convolutional neural networks. 2015 IEEE International Conference on Image Processing (ICIP). IEEE, 452–456.

Liu B, **Wei Y**, **Zhang Y**, **Yang Q. 2017**. Deep neural networks for high dimension, low sample size data. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), 2287–2293. Available from: https://www.ijcai.org/Proceedings/2017/0318.pdf

Lu YC, **Tung C**, **Kuo YF. 2020**. Identifying the species of harvested tuna and billfish using deep convolutional neural networks. *ICES Journal of Marine Science* **77**: 1318–1329.

MacLeod N, **Benfield M**, **Culverhouse P. 2010**. Time to automate identification. *Nature* **467**: 154–155.

Mäder P, **Boho D**, **Rzanny M**, **Seeland M**, **Wittich HC**, **Deggelmann A**, **Wäldchen J. 2021**. The flora incognita app–interactive plant species identification. *Methods in Ecology and Evolution* **12**: 1335–1342.

Marques ACR, **Raimundo M**, **Cavalheiro EMB**, **Salles LFP**, **Lyra C**, **Von Zuben FJ. 2018**. Ant genera identification using an ensemble of convolutional neural networks. *PLoS One* **13**: e0192011.

Miao Z, **Gaynor KM**, **Wang J**, **Liu Z**, **Muellerklein O**, **Norouzzadeh MS**, **McInturff A**, **Bowie RC**, **Nathan R**, **Yu SX**, **Getz WM. 2019**. Insights and approaches using deep learning to classify wildlife. *Scientific Reports* **9**: 1–9.

Milošević D, **Milosavljević A**, **Predić B**, **Medeiros AS**, **Savić-Zdravković D**, **Piperac MS**, **Kostić T**, **Spasić F**, **Leese F. 2020**. Application of deep learning in aquatic bioassessment: towards automated identification of non-biting midges. *Science of the Total Environment* **711**: 135160.

Nguyen H, **Maclagan SJ**, **Nguyen TD**, **Nguyen T**, **Flemons P**, **Andrews K**, **Ritchie EG**, **Phung D. 2017**. Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 40–49.

Norouzzadeh MS, **Nguyen A**, **Kosmala M**, **Swanson A**, **Palmer MS**, **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 USA* **115**: E5716–E5725.

Pinho C, **Harris DJ**, **Ferrand N. 2007**. Comparing patterns of nuclear and mitochondrial divergence in a cryptic species complex: the case of Iberian and North African wall lizards (*Podarcis*, Lacertidae). *Biological Journal of the Linnean Society* **91**: 121–133.

Pinho C, **Harris DJ**, **Ferrand N. 2008**. Non-equilibrium estimates of gene flow inferred from nuclear genealogies suggest that Iberian and North African wall lizards (*Podarcis* spp.) are an assemblage of incipient species. *BMC Evolutionary Biology* **8**: 63.

Rauf HT, **Lali MIU**, **Zahoor S**, **Shah SZH**, **Rehman AU**, **Bukhari SAC. 2019**. Visual features based automated identification of fish species using deep convolutional neural networks. *Computers and Electronics in Agriculture* **167**: 105075.

Renoult JP, **Geniez P**, **Bacquet P**, **Benoit L**, **Crochet PA. 2009**. Morphology and nuclear markers reveal extensive mitochondrial introgressions in the Iberian wall lizard species complex. *Molecular Ecology* **18**: 4298–4315.

Renoult JP, **Geniez P**, **Bacquet P**, **Guillaume CP**, **Crochet PA. 2010**. Systematics of the *Podarcis hispanicus*-complex (Sauria, Lacertidae) II: the valid name of the north-eastern Spanish form. *Zootaxa* **2500**: 58–68.

Salvi D, **Pinho C**, **Mendes J**, **Harris DJ. 2021**. Fossil-calibrated time tree of *Podarcis* wall lizards provides limited support for biogeographic calibration models. *Molecular Phylogenetics and Evolution* **161**: 107169.

dos Santos AA, **Goncalves WN. 2019**. Improving Pantanal fish species recognition through taxonomic ranks in convolutional neural networks. *Ecological Informatics* **53**: 100977.

Sá-Sousa P, **Pérez-Mellado V**, **Martínez-Solano I. 2009**. *Podarcis carbonelli*. The IUCN Red List of Threatened Species 2009: e.T61545A12512496. https://dx.doi.org/10.2305/IUCN.UK.2009.RLTS.T61545A12512496.en. Accessed on 21 September 2022.

Seeland M, **Rzanny M**, **Boho D**, **Wäldchen J**, **Mäder P. 2019**. Image-based classification of plant genus and family for trained and untrained plant species. *BMC Bioinformatics* **20**: 1–13.

Selvaraju RR, **Cogswell M**, **Das A**, **Vedantam R**, **Parikh D**, **Batra D. 2017**. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

Szegedy C, **Ioffe S**, **Vanhoucke V**, **Alemi A. 2017**. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence. **2018**: 4278–4284.

Szegedy C, **Vanhoucke V**, **Ioffe S**, **Shlens J**, **Wojna Z. 2015**. Rethinking the inception architecture for computer vision. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826.

The ImageMagick Development Team. 2021. ImageMagick. Retrieved from https://imagemagick.org

Wäldchen J, **Mäder P. 2018**. Machine learning for image based species identification. *Methods in Ecology and Evolution* **9**: 2216–2225.

Wilson EO. 2004. Taxonomy as a fundamental discipline (Godfray HCJ, Knapp S, eds.). *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **359**: 739–739.

Yang W, **Feiner N**, **Pinho C**, **While GM**, **Kaliontzopoulou A**, **Harris DJ**, **Salvi D**, **Uller T. 2021**. Extensive introgression and mosaic genomes of Mediterranean endemic lizards. *Nature Communications* **12**: 1–8.

Zachos FE. 2016. *Species concepts in biology*. Cham: Springer.

Zhou H, **Yan C**, **Huang H. 2016**. Tree species identification based on convolutional neural networks. 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). IEEE, 103–106.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

Table S1. Detailed classification results for male dorsal images in the two-class experiment.
Table S2. Detailed classification results for male head lateral images in the two-class experiment.
Table S3. Detailed classification results for female dorsal images in the two-class experiment.
Table S4. Detailed classification results for female head lateral images in the two-class experiment.
Table S5. Detailed classification results for male dorsal images in the nine-class experiment.
Table S6. Detailed classification results for male head lateral images in the nine-class experiment.
Table S7. Detailed classification results for female dorsal images in the nine-class experiment.
Table S8. Detailed classification results for female head lateral images in the nine-class experiment.