
USING DEEP LEARNING TO CLASSIFY IMAGES OF WALL
LIZARDS

Catarina Lopes Pinho

Thesis

Master in Modelling, Data Analytics and Decision Support Systems

Supervised by

João Gama

Carlos Ferreira

Antigoni Kaliontzopoulou

2021

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores por terem tornado este trabalho possível. Agradeço ao Professor João Gama pela confiança e pelo constante encorajamento e sugestões de melhoria. Ao Professor Carlos Ferreira agradeço as valiosas sugestões e a ajuda em algumas das partes mais complexas deste trabalho. À minha colega e amiga Antigoni Kaliontzopoulou, aqui tornada orientadora também, agradeço a cedência da maioria das imagens e a preciosa visão biológica do problema.

Ao Guilherme Caeiro-Dias agradeço o ter-me permitido utilizar dezenas de fotos da sua autoria.

Aos meus pais, Mário Jorge e Maria de Fátima Pinho, agradeço todo o apoio e encorajamento que sempre me deram em todas as circunstâncias, incluindo na decisão algo louca de começar, aos 40 anos, um mestrado numa área diferente da minha área de especialização.

Ao meu companheiro Daniel Cunha agradeço a paciência infinita e o apoio incondicional, mesmo nas muitas noites em que teve de “aguentar o barco” sozinho. Agradeço ainda em particular a ajuda na utilização do Photoshop, sem a qual esta tese teria sido muito mais complicada.

Às minhas filhotas Rita, Inês e Sara agradeço o carinho e compreensão que me ajudaram a ultrapassar as etapas mais difíceis deste percurso e a paciência com que lidaram com as minhas ausências. Tudo é mais fácil com o vosso amor!

Finalmente, dedico esta tese à minha avó Rosa Ângela, desaparecida durante a reta final da escrita desta dissertação, agradecendo-lhe por tudo o que me deu: o carinho incomensurável, o exemplo de resiliência e perseverança, e a constante lembrança de que pouco mais importa na vida do que sermos e fazermos os outros felizes.

Abstract

Identifying species is an important task in biology, medicine, pharmacology, agriculture and biodiversity conservation. Modern taxonomy, the discipline devoted to the description and identification of species, faces a decrease in its workforce while dealing with the challenge of describing the Earth's vanishing biodiversity. In this context, automated identification of species becomes of major importance. Automated image identification is a thriving field of machine learning, with deep learning algorithms based on convolutional neural networks revolutionizing the field in recent years. These methods are also taking their first steps in the identification of images of biological taxa in a variety of different contexts and taxonomic scopes.

In this work, we took a deep learning approach to classify images of Iberian and North African wall lizards, a group of cryptic species in which identification requires expert intervention. We addressed two problems: 1) a two-class problem focusing on the distinction between two species, *Podarvis bocagei* and *Podarvis lusitanicus* and 2) a nine-class problem involving all the species currently described in the group. Three different deep learning architectures were tested in both cases. In the two-class problem, classification success was high, reaching as high as 97.1% and 95.9% for ensemble models applied to male and female lizards, respectively. Classification in the nine-class problem was not as successful, highlighting the difficulties inherent to this group of cryptic species. However, results improved when predictions from different perspectives were combined, reaching 95.3% and 89.7% for males and females, respectively. These results suggest the utility of deep learning algorithms in the identification of cryptic species, providing promising resources in the taxonomical, evolutionary and conservation research.

Sumário

A identificação de espécies é uma tarefa importante na biologia, medicina, farmacologia, agricultura e conservação da biodiversidade. A taxonomia, disciplina que se dedica à identificação e classificação das espécies, enfrenta atualmente uma diminuição significativa do número de profissionais, enquanto abraça a tarefa de descrever a biodiversidade da Terra, altamente ameaçada. Neste contexto, a identificação automática de espécies torna-se particularmente importante. A identificação automática de imagens é uma área científica em expansão, com os algoritmos de *deep learning* baseados em redes neurais de convolução a gerar uma revolução neste campo nos últimos anos. Estes métodos têm também dado os seus primeiros passos na identificação de imagens de espécies biológicas em diversos contextos.

Neste trabalho usou-se uma abordagem de *deep learning* para classificar imagens de lagartixas do género *Podarvis* da Península Ibérica e do Norte de África, um grupo de espécies crípticas cuja identificação requer a intervenção de peritos. Foram abordados dois problemas: 1) um problema de classificação binária focado na distinção entre *Podarvis bocagei* e *Podarvis lusitanicus*; 2) um problema com nove classes envolvendo todas as espécies atualmente descritas neste grupo. Testaram-se três arquiteturas de *deep learning* diferentes em ambos os casos. No problema binário, o sucesso de classificação obtido foi alto, alcançando os 97.1% e 95.9% no caso de combinações de modelos aplicados a imagens de machos e fêmeas, respetivamente. No problema com nove classes a classificação não foi tão bem sucedida, enfatizando as dificuldades inerentes à identificação de espécies crípticas, particularmente para casos como *P. liolepis*. No entanto, os resultados melhoraram consideravelmente combinando previsões obtidas da análise de diferentes modelos e perspectivas, chegando aos 95.3% e 89.7% para machos e fêmeas, respetivamente. Estes resultados sugerem a utilidade dos algoritmos de *deep learning* na identificação de espécies crípticas, constituindo recursos promissores na investigação em taxonomia, evolução e conservação.

Contents

1. Introduction	1
1.1 Problem and motivation	1
1.2 Organization of the thesis	2
2. Literature Review	3
2.1 Identifying biological species	3
2.1.1 The species problem.....	3
2.1.2 The challenges of taxonomy	4
2.2 Automated Image Classification.....	5
2.2.1 A historical perspective.....	5
2.2.2 A brief survey on Convolutional Neural Networks	6
2.2.3 Applications of deep learning algorithms in species identification.....	12
3. Methodology	16
3.1 The model system: Iberian and North Africa wall lizards	16
3.2 Description of image data.....	17
3.3 Image pre-processing	18
3.4 Analytical procedures.....	19
3.4.1 Distinction between <i>P. bocagei</i> and <i>P. lusitanicus</i>	19
3.4.2 Distinction between the nine species in the Iberian and North African clade.....	21
4. Results	23
4.1 Distinction between <i>P. bocagei</i> and <i>P. lusitanicus</i>	23
4.2 Distinction between all nine species in the Iberian and North African clade.....	26
5. Discussion	37
6. Conclusions and future perspectives	43
References	44
Appendix 1. Detailed evaluation results	51
A1.1 Two-classes case	51
A1.2 Nine-classes case	54

List of figures

Figure 1. The basic architecture of a CNN	7
Figure 2. Visual example of the convolution operation.....	8
Figure 3. Example of a dorsal (left) and head lateral view (right) for the same individual.	17
Figure 4. Example of Grad-CAM heatmaps obtained for <i>Podarcis lusitanicus</i>	26
Figure 5. Classification success for male dorsal images based on InceptionResNetV2 results.....	29
Figure 6. Classification success for male head lateral images based on ResNet50 results.	30
Figure 7. Classification success of female dorsal images based on ResNet50 results.....	31
Figure 8. Classification success for female head lateral images based on Inception ResNet V2 results.....	32
Figure 9. Confusion matrix for male (upper) and female (lower) image classification based on a combination of predictions from the six models applied.....	34

List of tables

Table 1. A summary of relevant articles using deep learning on biological images	14
Table 2. Number of images per class, sex and view.	18
Table 3. Evaluation of the three tested architectures in the four datasets for the two-class problem.	24
Table 4. Assessing the utility of ensemble models for image classification in the two-class problem.	25
Table 5. Evaluation of the three tested architectures in the four datasets for the nine-class problem.....	27
Table 6. Assessing the utility of ensemble models for image classification in the nine-class problem	28
Table 7. Summary of Grad-CAM results for each class (males).....	35
Table 8. Summary of Grad-CAM results for each class (females).....	36

1. Introduction

This chapter consists of a brief description of the problem addressed in this thesis and of an overview of the thesis' organization.

1.1 Problem and motivation

This work bridges two very different research areas: biology and computer science. Automated image classification is an active research field in computer science, and its developments have fuelled diverse applications in other scientific areas, including biology. In the biological sciences, however, automated image classification is far from being developed at its full potential, often because of the multidisciplinary nature of the task, which imposes challenges for a biologist with a regular training. Biological image automated classification can be especially helpful in two contexts, which are not mutually exclusive: i) one in which the data set is so large that processing images and making the identification by humans is very time-consuming and ii) one in which identification is particularly difficult and relying on expert knowledge or alternative techniques such as DNA sequencing, which are not practical or affordable at a large scale.

This work falls mainly within the second context. The problem that will be tackled by this thesis is the automated identification of species of wall lizards (genus *Podarvis*) from the Iberian Peninsula and North Africa, based on a comprehensive image data set and using deep learning techniques. Wall lizards are a group for which traditional classification methods often fail and in which expert intervention and/or (most often) laboratory DNA sequencing are required for species identification. This is the first time that automated image classification techniques have been applied in this system.

The motivation for this work is two-fold:

1 – for practical purposes, identifying the species from the group analysed without the need for expert intervention will have strong positive repercussions in both subsequent research and monitoring (particularly of endangered species).

2 – in a more conceptual context, identifying what humans have so far been unable to clearly point out in this system (the features that distinguish species) is important in the study of how these and other species evolve.

1.2 Organization of the thesis

This thesis is organized in five chapters. In this first chapter, I provide a brief introduction to the topic and the thesis. Chapter two is dedicated to a review of the pertinent literature, focusing both on a biological scope, namely the definition of species and the challenges posed to taxonomists in present days, and a brief revision of the literature on automated image classification and convolutional neural networks, the tools of choice for this purpose in recent years. Chapter two finishes with a revision of studies where deep learning methods have been applied for the identification of biological species. In chapter three I present the biological system used and explain the methodological procedures adopted to tackle both classification problems addressed in this thesis. In chapter four I present the main results obtained. Chapter five consists of a discussion of the results obtained in the light of the current knowledge both on the methodologies applied and on the biological system under investigation. Finally, chapter six provides a brief conclusion about the work and offers future perspectives.

2. Literature Review

This chapter provides a brief summary of the literature in the two main topics of this project: the identification of species, on one hand, and deep learning algorithms, on the other. A third section provides a bridge between these two main fields of knowledge, summarizing available studies applying deep learning tools for the identification of biological images.

2.1 Identifying biological species

2.1.1 The species problem

The hierarchical Linnaean system is the framework used by biologists for the classification of life forms. It constitutes in grouping organisms in smaller and smaller groups, which are included within each other, to provide a hierarchical categorization of living diversity. Most levels in this classification system are artificial (that is, they are human constructions meant for categorizing and not meaningful natural entities). The only exception is perhaps the species, which is the basal category of Linnaean taxonomy. Despite the fact that the topic of the reality of species itself has been often debated in the scientific literature (see Hey et al., (2003) and references therein), most biologists today agree that species exist independently of human observers and that they correspond to real natural discontinuities.

However, objectively defining and identifying species is a very difficult task, and the set of questions around this topic has been called *the species problem* in the evolutionary biology literature (Hey, 2001; Zachos, 2016). It is virtually impossible to use a single species definition criterion (usually called “species concept”) that is applicable to all life forms and situations. A recent review identifies about 30 different species concepts (Zachos, 2016), although many of them share the main idea of species as independently evolving lineages. Part of the difficulty comes from trying to impose a discrete classification on the outcomes

of a continuous evolutionary process. Recent views thus separate the problems of defining species, on one hand, and of practically delimiting them, on the other. Whereas the first is a conceptual problem, the second is a methodological one, and named species can be thought of as approximations of the real evolutionary entities (Ghiselin, 2001; Zachos, 2016).

2.1.2 The challenges of taxonomy

Despite the conceptual difficulties associated to the definition of species, naming species is a task of utmost importance, and taxonomy – the field of biology concerned with the classification and naming of life forms – is considered as a fundamental discipline (Wilson, 2004). First, biological research requires the use of a common system that can be used across all disciplines and between different researchers. Second, biological names are important in a variety of different contexts outside biological research: medicine, pharmacology, agriculture, in museum collections, in international trade regulations, and biodiversity monitoring and conservation.

Taxonomy deals with the task of delimiting, identifying, and characterizing species. Traditionally this was done exclusively using comparative morphology or the study of other phenotypic characters (like calls in insects, birds, or amphibians). For the past decades, however, the use of molecular analyses (e.g. DNA sequencing) has opened new perspectives in the field, particularly, but not exclusively, in the case of cryptic species (that is, species that are morphologically very similar; Bickford et al., 2007). Recent approaches involve the so-called “integrative taxonomy” (Padial et al., 2010), which bridges different disciplines and techniques, including not only morphology and molecular investigations but also analyses of ecology analyses and reproductive isolation.

Current taxonomic research faces several important challenges: first, the acknowledgement that the Earth’s biodiversity is far from completely described (it is estimated that less than 20% of the world species have been described so far). Second, the fact that the Earth is facing its sixth (and largely human-induced) mass extinction (Ceballos et al., 2015) and that correctly cataloguing and providing tools for species identification is critical for the monitoring and preservation of biodiversity. Third, the well-known decrease of the

taxonomic workforce over the past decades, a problem known as the *taxonomic impediment* (Drew, 2011; Hopkins & Freckleton, 2002). One of the biggest constraints to the conservation of biodiversity is the lack of knowledge on species' distribution ranges and their modifications over time, which in turn is related to a lack of experts to perform species identification. In this context, the need for automatic species identification has been stressed by different authors, and it is finally moving forward at a fast pace (see section 2.2.3) (Gaston & O'Neill, 2004; MacLeod et al., 2010).

2.2 Automated Image Classification

2.2.1 A historical perspective

Image Classification can be defined as the task of labelling images based on predefined classes (Rawat & Wang, 2017). It has been used for a wide variety of applications and it also constitutes the backbone of other computer vision tasks, such as object detection, localization, or segmentation (Karpathy, 2016). Although image classification is trivial for humans, it is a highly challenging task for a machine, since it requires generalizing well over the diversity typically exhibited within a class. In addition, there can be variations in perspective, scale or illumination, the background, partially covered or deformed objects, etc. (Ciresan et al., 2011; Sejnowski, 2018).

For many years, the standard approach for image classification was a dual-stage procedure (Rawat & Wang, 2017): in a first stage, relevant features were extracted into descriptors, in a process involving a variable degree of human intervention; in a second stage, a classifier was applied to the features extracted. Feature extraction was highly laborious and time-consuming, and often specific to each project (LeCun et al., 1998).

The development of deep learning algorithms brought a major change in the field. *Deep learning* is a field of machine learning that focuses on learning high-level abstractions from data, circumventing the need for hand-designed feature extraction (Goodfellow et al., 2016; Guo et al., 2016). These methods experienced a major surge around 10 years ago, with the

onset of the so-called “deep learning revolution”. Deep learning algorithms were found to be suitable for handling big data with successful applications in diverse areas (Liu et al., 2017). A turning point in Image Classification was when AlexNet (Krizhevsky et al. 2012) won the prestigious ImageNet Large Scale Visual Recognition Challenge (ILSVRC; Russakovsky et al., 2015). The object of this yearly competition is the ImageNet data set (Deng et al., 2009) which is a set of millions of annotated images developed for academic purposes.

2.2.2 A brief survey on Convolutional Neural Networks

Among the various deep learning methods (for a summary see (Guo et al., 2016) and (Liu et al., 2017)), those that have been more widely used for Image Classification are, without a doubt, Convolutional Neural Networks (CNNs). Neural networks are a type of machine learning algorithms that mimic the functioning of the animal brains. The next paragraphs briefly review the most important aspects of CNNs (for a more complete overview see e.g. Rawat & Wang, 2017).

2.2.2.1 The basic architecture of a CNN

CNNs are a type of neural network that processes data with a grid-like topology, such as images (Goodfellow et al., 2016). There are three major types of layers in a CNN, each of which plays a different role in the hierarchical structure: convolution layers, pooling layers (also called “subsampling layers”) and fully connected layers (see Figure 1).

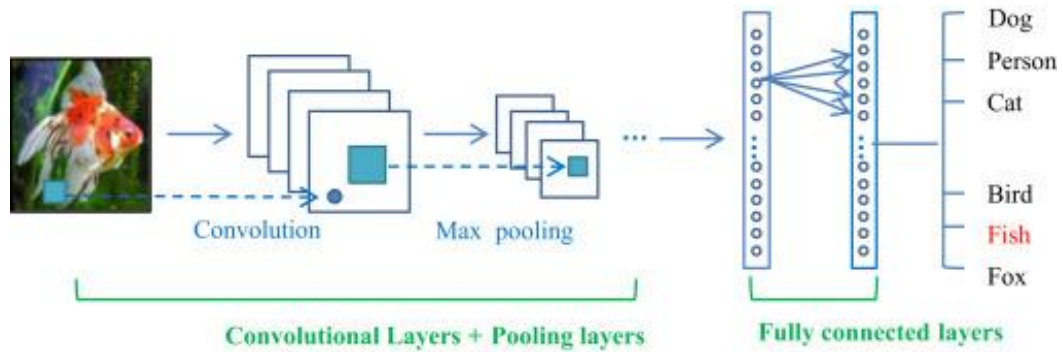


Figure 1. The basic architecture of a CNN (from Guo et al., 2016).

Convolution layers perform a convolution operation on the input using filter matrices (also called kernels). The convolution is an operation on two functions such that (in its discrete form):

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{+\infty} x(a)w(t - a)$$

When dealing with two-dimensional inputs such as images (I) and two-dimensional kernels (K), this operation can be written as (Goodfellow et al., 2016):

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, k - n)$$

The convolution thus performs a sliding weighted sum by moving the filter over the image, as shown in Figure 2:

$$\begin{pmatrix}
 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
 1 & 1 & 0 & 0 & 0 & 0 & 0
 \end{pmatrix}
 *
 \begin{pmatrix}
 1 & 0 & 1 \\
 0 & 1 & 0 \\
 1 & 0 & 1
 \end{pmatrix}
 =
 \begin{pmatrix}
 1 & 4 & 3 & 4 & 1 \\
 1 & 2 & 4 & 3 & 3 \\
 1 & 2 & 3 & 4 & 1 \\
 1 & 3 & 3 & 1 & 1 \\
 3 & 3 & 1 & 1 & 0
 \end{pmatrix}$$

I
 K
 $I * K$

Figure 2. Visual example of the convolution operation (from <https://i.stack.imgur.com/RhEhb.png>)

The resulting matrix is called a feature map. This operation has very convenient properties: first, the process of sliding the filter along the image allows the algorithm to learn correlations among neighbouring pixels, which helps detecting interesting features in images (e.g. corners or edges). Second, this operation grants invariance to the specific location of the object. And finally, because the kernels are in general smaller than the input, there is a high degree of parameter sharing between different neurons, which makes the process more efficient than a typical neural network layer (Goodfellow et al., 2016; Guo et al., 2016).

Usually, after a convolution layer, which includes only linear transformations, there is a non-linear activation stage (Goodfellow et al., 2016). The goal of this step is to make the process more generalizable to the (usually non-linear) nature of the data. Although various types of non-linear transformations have been proposed, the standard in machine learning applications is the Rectified Linear Unit (also called ReLU) activation (Nair & Hinton, 2010; Zeiler et al., 2013). This transformation replaces all negative values by zero (while maintaining all non-negative values the same).

Pooling layers subsample feature maps by replacing each group of neighbouring neurons by a summary statistic (Goodfellow et al., 2016). This reduces the spatial resolution of the feature map, making the network more manageable, preventing overfitting and making the

process more robust to noise (Guo et al., 2016; Liu et al., 2017; Rawat & Wang, 2017). Max pooling has become the standard in the field (Ciresan et al., 2011; Guo et al., 2016), although other kinds (e.g. average pooling) are also used.

Finally, *fully connected layers* are typically the last in a CNN architecture and function as traditional neural networks. They are heavily parameterized (Guo et al., 2016). These layers convert two-dimensional feature maps into one-dimensional vectors and act as the classifier. For classification problems, the *softmax* activation is generally used (Rawat & Wang, 2017). The softmax is the generalization of the logistic regression classifier to a multiclass problem and is used to normalize the output of the previous layer to a probability distribution over n different classes (Goodfellow et al., 2016; Karpathy 2016).

2.2.2.2 Training a CNN

Training a CNN is a time-consuming and difficult task. As in other kinds of neural networks, training involves initialization of parameter values (filters and weights), propagating the information in a feed-forward manner through the various layers, obtaining predictions, calculation of the gradient of the error with respect to the parameters by backpropagation, and updating the parameters based on gradient descent; the process is then repeated for all images in the training set. Because of the very large number of parameters involved, typically in the order of millions, CNNs are prone to overfitting and require very large sample sizes, which are not always available. Therefore, various strategies have been developed to overcome problems associated with training:

- *Data augmentation* artificially increases the number of images and creates noise in the data set by producing slight changes in the original images. Several techniques have been used for this purpose, such as shifting, rotating, zooming, cropping or flipping the image (e.g. Gómez-Ríos et al., 2019), or altering the intensities of the RGB channels (Krizhevsky et al. 2012), for example.
- *Dropout* (Hinton et al., 2012) is a widely used technique that relies on stochastically deleting part of the neurons in each training example to prevent the co-adaptation of feature detectors and hence improve the generalization ability of the model (Baldi & Sadowski, 2013).

- *Pre-training* refers to the initialization of the training using weights and filters obtained from training the network using a different data set. This process, also called *transfer learning*, has been shown to accelerate the learning process and decrease overfitting (Guo et al., 2016). For example, various architectures pre-trained from the ImageNet data set are freely available and are often used for different tasks. Typically, all layers of the model are initialized with the parameters obtained from pre-training, except the last fully connected layers, which are specific for the classification problem being addressed. Fine tuning of the parameters with respect to the specific task is then carried out.

2.2.2.3 Examples of successful architectures

Different architectures have been proposed to deal with the problem of image classification. The first CNN was LeNet (LeCun et al., 1998), a simple network consisting of 7 layers (with few channels each) but already including the fundamental aspects mentioned above. AlexNet (Krizhevsky et al. 2012), the first CNN winning the ILSVRC, was much more complex, involving 5 convolution layers and kernels with 192 channels, and 60 million parameters overall (Russakovsky et al., 2015). This complexity was handled using techniques to avoid overfitting, such as data augmentation and dropout. In 2014, two of the competitors to the ILSVRC brought important breakthroughs: VGG (Simonyan & Zisserman, 2014) and GoogLeNet (Szegedy et al., 2014). VGG shows two improvements from AlexNet: the network is much deeper, but the filters are smaller. GoogLeNet (the winner of the ILSVRC2014) takes the increase in the depth of the network even further, while decreasing the number of parameters via the use of a module called Inception, which applies innovative techniques (like a 1x1 convolution and the concatenation of simultaneous operations) to reduce the dimensionality of the problem while achieving very good accuracy. Inception V3 (Szegedy et al., 2015), another widely used architecture, is built on improvements of the Inception module. Another architecture that represented a major innovation in the field was ResNet (He et al., 2016), which won the ILSVRC in 2015. This introduced 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. Xception (Chollet, 2017), inspired in Inception, introduced the concept of depthwise convolution. Inception-

ResNet-V2 (Szegedy et al., 2017) combines the Inception architecture with residual learning, showing significant improvements both in training speed and in classification success. DenseNet (Huang et al., 2018) is another architecture that extends the idea of residual learning even further, creating connections between all the layers. Finally, another architecture worth mentioning is MobileNet (Howard et al., 2017), a lighter yet efficient architecture specific for resource-constrained environments such as mobile device processors.

2.2.2.4 Deep learning explainers

Deep learning algorithms are often considered black boxes; there is a trade-off between accuracy and explainability, and the deeper the algorithm the more obscure it is: high level features extracted from the models are typically not traceable or interpretable (Buhrmester et al., 2019). Consequently, a model may appear to be working but could be looking at completely irrelevant characters. For example, a classifier that was built to distinguish wolf and dog images was actually looking at the snow in the background to make predictions (Ribeiro et al., 2016); evidently, visualizing what a classifier is using is important to validate the results of the model before it is deployed and, in the case of classifiers involving human subjects, it is also an ethical imperative (Buhrmester et al., 2019). Several different explainers have been proposed in the field of Computer Vision (for a review see Buhrmester et al., 2019). Many of the proposed methods detect important pixels by analysing the effects of changing their intensity on the prediction (e.g Zeiler & Fergus, 2014). A growingly popular method, particularly in the case of biological images (see section 3.2.3), is Grad-CAM (Selvaraju et al., 2017, 2020) – Gradient-Weighted Class Activation Mapping. This method 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.

2.2.3 Applications of deep learning algorithms in species identification

As in other fields of knowledge, the easiness of obtaining digital photos has led to a huge increase of biological images as a potential source of data. This happens at a time when the number of taxonomic experts is decreasing and where monitoring biodiversity is more urgent than ever (see section 2.1.2; Wäldchen & Mäder, 2018). Although automated biological taxa identification is not yet a common task, there is a growing body of literature with examples of successful applications, particularly in the last two years (see Table 1). Several large scale identification tools are becoming more frequent (e.g. Barré et al., 2017, Buschbacher et al., 2020), as well as freely available mobile applications for the general public (e.g. iNaturalist Seek (https://www.inaturalist.org/pages/seek_app), Flora Incognita (Mäder et al., 2021), Pl@ntNet (Affouard et al., 2017), or even non-specialized apps like Google Lens) which are changing biodiversity monitoring (Bonnet et al., 2020). In recent years, competitions aiming at image identification have also been promoted, such as LifeCLEF (www.imageclef.org) or the iNaturalist challenge (Van Horn et al., 2018; <https://www.kaggle.com/c/inaturalist-2018>).

Wäldchen et al., (2018) point out some of the challenges of applying deep learning techniques to biological images: i) typically very high number of classes; ii) large intraspecific visual variation; iii) low interspecific variation; iv) the possibility of dealing with untrained taxa; v) the variation induced by the image acquisition process. The first problem can be handled by reducing the taxonomic scope, and the last can be dealt with by using specific protocols for image acquisition (e.g. Milošević et al., 2020). However, the other problems are common across different contexts and mostly unavoidable.

One of the most challenging types of data set in biological image identification is that generated by camera traps (motion-activated cameras aimed at capturing secretive wildlife and that generate large quantities of very diverse images). One of the first studies using deep learning was precisely directed at this type of data set (Chen et al., 2014). Although that study in particular did not achieve a good classification accuracy (only around 38%), it still performed better than benchmark methods used at the time. Other studies of this type followed (e.g. Miao et al., 2019; Nguyen et al., 2017; Norouzzadeh et al., 2018). In particular, Norouzzadeh et al. (2018) performed a thorough study involving different architectures and approaches, obtaining high accuracy (maximum of 93.8% using

ResNet50) and showing that deep learning methods are a powerful tool that may alleviate the human burden of annotating images.

Interestingly, very few of the studies analysed use or report pre-processing techniques, and only one, focusing on the very specific case of underwater images (Gómez-Ríos et al., 2019b), compares the effect of these techniques on the classification outcomes. This study found that deblurring, and contrast and brightness enhancement (CBE) slightly improve accuracy, while saliency decreases it.

In general, and as expected, the data sets with the higher number of classes are those that generate a worse predictive ability (e.g. Hansen et al., 2020; Picek et al., 2020; Seeland et al., 2019). To cope with relatively small sample sizes, many authors opt by using data augmentation techniques, which vary depending on the study. Curiously, the two studies performing a comparison of the utility of data augmentation do not agree: Hsiang et al., (2019) obtained worst results when using augmenting techniques whereas Gómez-Ríos et al., (2019b) report improvements in accuracy. Another technique that is almost standard is the use of transfer learning, usually by using parameters pre-trained from ImageNet. Again, one study found this option not to bring any advantage (Norouzzadeh et al., 2018) whereas another showed consistently improved results (Buschbacher et al., 2020).

One interesting study in the context of this work, because of similarities in the dimension of the dataset and the sharing of similar goals, is Milošević et al., (2020). This study includes 1846 images of midge larvae comprising 10 different species. The authors used a standard image acquisition protocol and ResNet50, with data augmentation and transfer learning, and obtained excellent results (99.5% at the species level). The authors used GradCAM (Selvaraju et al., 2020). This is the most popular explainer used in the reviewed literature; in fact, five other studies use the same method (Banan et al., 2020; dos Santos & Goncalves, 2019; Lu et al., 2020a; Miao et al., 2019; Seeland et al., 2019a), albeit deconvolutional networks and saliency maps were also used (by Lee et al., (2015) and Lu et al., (2020), respectively).

It should be noted that some of the studies reviewed only report validation accuracy, which is optimized during learning, and therefore success cannot be straightforwardly compared to those that report testing accuracy.

Nevertheless, although some reporting bias could exist, in summary most studies appear to be mostly successful (see reported accuracies in Table 1).

Table 1. A summary of relevant articles using deep learning on biological images

Reference	Taxa	Types of images	Number of images	Number of classes	Improvements	Accuracy	Architecture
Chen et al., (2014)	Animals	camera trap	23876	20	DA	38.3	specific
Lee et al., (2015)	Plants	standard	2816	44	DA, TLIN	98.1- 99.5	AlexNet
Zhou et al., (2016)	Trees	standard	2358	25		87-91	LeNet
Barré et al., (2017)	Trees	mixed	26624/6000/1526	184/60/32	DA	86.3-97.9	specific
Gogul & Kumar, (2017)	Plants	field	8189/2240	102/28	TLIN	92-93	Inception v3, Xception, Overfeat
Nguyen et al., (2017)	Animals	camera trap	30000/44536	3/6	TLIN	85-90	Lite AlexNet, VGG16, ResNet50
Marques et al., (2018)	Ants (genera)	standard	44806	57	TLIN	75-98	AlexNet
Norouzzadeh et al., (2018)	Animals	camera trap	301400	48	DA, TLIN (not helpful)	93.8	AlexNet, NiN, VGG22, GoogLeNet, ResNet18,34,50,101,152
Arzar et al., (2019)	Butterflies (genera)	Field	120	4	TLIN?	97.5	GoogLeNet
Gómez-Ríos et al., (2019a)	Coral	Field	766/1123	14/8	DA, TLIN	98	Inception ResNet50,152 DenseNet121,161
Gómez-Ríos et al., (2019b)	Coral	Field	409	14	DA, TLIN	85-93	Inception, ResNet50,152, DenseNet121,161
Hsiang et al., (2019)	Foraminifera	Standard	34640	36	DA (not helpful)	87.4	VGG16 DenseNet121 Inception
Miao et al., (2019)	Mammals	camera trap	111467	20		83	VGG16 ResNet50
Rauf et al., (2019)	Cyprinids (fish)	Standard	438 (total, divided into 3 body parts)	6		84-96	AlexNet LeNet GoogLeNet ResNet50 VGG16 VGG32 modified
dos Santos & Gonçalves, (2019)	Fish	Mixed	~9700	68	DA, TLIN	~88	Inception, others
Seeland et al., (2019)	Plants	Field	117713	1000 species, 516 genera, 124 families	DA, TLIN	82.2-88.4	Inception-ResNet-v2

Table 1. (cont.)

Reference	Taxa	Types of images	Number of images	Number of classes	Improvements	Accuracy	Architecture
Almryad & Kutucu, (2020)	Butterflies (genera)	Field	9000	10	TLIN	79.5	VGG16,19 ResNet50
Banan et al., (2020)	Carps	Standard	409	4	DA, TLIN	100	VGG16
Buschbacher et al., (2020)	Bees	Standard	7595	124	DA, CW, TLIN	94	MobileNetV2
Hansen et al., (2020)	Carabidae (Beetles)	Standard	63364	291	TLIN	51.9-74.9	Inception v3
Lu et al., (2020)	Fish	Standard	16517	10	DA, TLIN	75-98	VGG16
Miele et al., (2020)	Muridae (mice)	Standard	1500	3	TL, modern to fossil	~100 except fossils	ResNet50-V2
Milošević et al., (2020)	Midges (insects)	Standard	1846	10	DA, TLIN	99.5	ResNet50
Picek et al., (2020)	Snakes	Mixed	~290000	783		<63 (F1)	ResNet50-V2
Raphael et al., (2020)	Corals	Field	5000	11	DA, TLIN	80.1	VGG16

Notes: DA, data augmentation; TLIN, transfer learning from ImageNet; TL, transfer learning (not from ImageNet), CW, class weighting. Accuracy is shown as a % (note one case where F1 was reported instead)

3. Methodology

This chapter describes the methods adopted in this work, starting with a brief introduction to the model system, a description of the data set, and then more particular aspects of the workflow.

3.1 The model system: Iberian and North Africa wall lizards

Wall lizards (genus *Podarcis*) are a group of small diurnal lizards that have a circum-Mediterranean distribution. Currently the genus comprises 25 species (www.reptile-database.org), although taxonomic revisions are ongoing (e.g. Caeiro-Dias et al., 2021) and this number is clearly an underestimate. Within this genus, several evolutionary and geographic coherent groups have been identified: this is the case of the Iberian and North African group, also known as the *Podarcis hispanicus* complex, a clade of several closely-related species that started to diverge around 7.5 million years ago (Salvi et al., 2021).

The species belonging to the Iberian and North African group are overall very similar morphologically, and the large intraspecific diversity overwhelms interspecific differences (Kaliontzopoulou, et al., 2012b). Because of this difficulty, external morphology is usually not considered a reliable indicator to distinguish species. The identification of evolutionary units, the description of species and the assessment of distribution maps in this group have been based mostly on genetics and/or conducted by highly experienced observers (Caeiro-Dias et al., 2018; Geniez et al., 2007; Kaliontzopoulou et al., 2012b; Renoult et al., 2010). Because these lizards are frequently used as models for evolutionary, ecological, physiological, and parasitological studies, and because at least one species (*P. carbonelli*) is endangered and requiring careful distribution monitoring, the need for genetic analysis or a taxonomic expert identification is an urgent necessity.

Nowadays, seven species are recognized in international databases: *P. bocagei*, *P. carbonelli*, *P. guadarramae*, *P. hispanicus*, *P. liolepis*, *P. vaucheri* and *P. virescens*; two more will be added in the near future (*P. lusitanicus*; Caeiro-Dias et al., 2021) and *P. tunesiacus* (Faria et al. in preparation). Further candidate species may exist inside some of the taxa, although these will not be analysed in the present study.

3.2 Description of image data

The images that have been used as input for this work were obtained between 2005 and 2015 in the scope of two PhD theses (Caeiro-Dias, 2018; Kaliontzopoulou, 2010). These images are as standardized as possible (e.g. the same perspectives were generally 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. Part of the images were used for geometric morphometric analyses as well as for obtaining scale count data (e.g. (Kaliontzopoulou, et al., 2012b), but so far had not been used for a study of automatic classification. We used images from two perspectives: a dorsal view focusing on the whole body (the tail, which the lizards often autotomize, and which may or may not appear in the image) and a lateral close-up of the head. Figure 3 represents examples of these two perspectives (before pre-processing) for the same individual. The sets also include both males and females. This genus exhibits marked sexual dimorphism, that is, males and females are frequently morphologically distinct, hence we considered each sex separately in the analyses. That is, for each problem addressed (see below), we analysed four datasets separately, corresponding to dorsal and head lateral images, and males and females for each perspective.

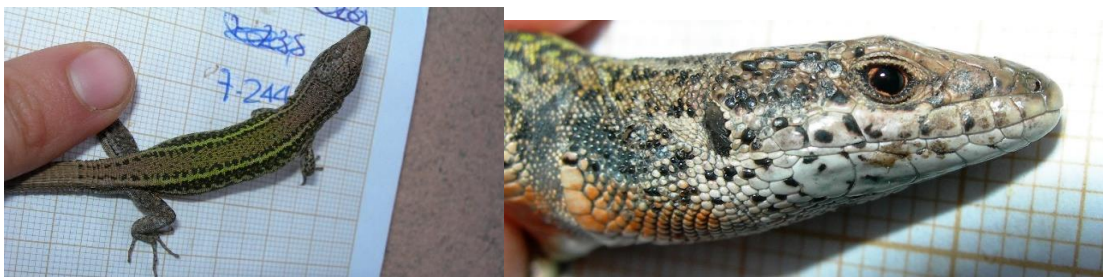


Figure 3. Example of a dorsal (left) and head lateral view (right) for the same individual.

The sets include individuals from the nine classes mentioned in the previous section, coming from different collection localities (ranging from 4 to 21 localities per class, with

this number higher in species with a larger geographic distribution). The “locality” label was ignored in this work, but it adds to the natural variability of the data examined.

Table 2 shows the composition of the data set.

Table 2. Number of images per class, sex and view.

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

3.3 Image pre-processing

The vast majority of the original images were already similarly oriented (that is, snouts pointing to the right), so we chose to maintain this feature to reduce complexity. Therefore, the first pre-processing step involved rotating or horizontally flipping the few images not conforming to this trend. After this, images were centered, cropped, converted to square format and resized to the same dimensions using the ImageMagick 7.0.10 software (The ImageMagick Development Team, 2021). This process was fairly automatic, but images were carefully checked (and manually corrected when needed). Although background removal is not mandatory for this type of analysis, many images included hand-written labels in the background which were likely to influence classification outcomes (e.g. different numbers were used for different species). Instead of manually manipulating

individual images to remove such labels, we opted by removing 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.

3.4 Analytical procedures

We addressed two main problems in this thesis: 1) a simple two-class problem: the distinction between two of the target species, *P. bocagei* and *P. lusitanicus*; and 2) a more complex nine-class problem: distinguishing all the recognized species in the Iberian clade. Because the methods for each problem were slightly different (although the same overall framework was used), we detail each workflow separately. All representative scripts have been deposited in Github (<https://github.com/catpinho>).

3.4.1 Distinction between *P. bocagei* and *P. lusitanicus*

This species pair was chosen for this initial analysis because the two species occur together, often in the same walls, throughout the northwest of the Iberian Peninsula. Hence, distinguishing between them is often of practical interest since the place of collection cannot help in this case (as it often does with other species). Moreover, although generally similar, as any other species in the genus, these two in particular show important differences in size, coloration and head shape (Gomes et al., 2016; Kaliontzopoulou, et al., 2012a), which makes them good candidates to assess the usefulness of computer vision models.

All procedures described below were conducted in the exact same manner for all four datasets. Prior to analyses, datasets (including 244 to 313 individual photographs from both species) were divided into five folds of the same size, maintaining class frequencies, based on which we created the data sets 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 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 x 224 pixels. We used data augmentation in the training datasets since initial experiments suggested it greatly reduced overfitting. Besides rescaling the data so that all values fall between 0 and 1 (which is a common procedure also to validation and testing datasets), data augmentation parameters 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. We did not augment data by flipping images since our dataset did not vary in this respect (see comment above about image orientation).

We chose three architectures for this work based on the literature review: InceptionV3, ResNet50 and InceptionResNetV2, all of which directly available in Keras. We initialized the models with weights pre-trained from ImageNet, a common practice in the field. 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 of a single unit using the sigmoid activation (for our binary classification problem). For all cases training was carried out using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001 (although in preliminary tests we experimented with different learning rates). A binary cross-entropy loss function was used. The learning process was conducted for 1000 epochs and with a batch size of 32. Because the classes in our datasets are unbalanced and preliminary runs showed an advantage in this procedure, class weights were used during model fitting to ensure that the lower frequency class (*P. lusitanicus*) receives more “attention” of the algorithm during the learning process. This was ensured by using the “balanced” heuristic in python module scikit-learn’s `compute_class_weights` function and providing resulting weights during training. The classification success of validation data set for each cross-validation replicate was monitored during the learning stage. After training, learning curves for both the training and validation data sets were inspected using TensorFlow visualization toolkit TensorBoard.

The models obtained at the end of the 1000 epochs were used to make predictions and evaluate the performance of the methods on each test set. A vector of probabilistic predictions for the whole data set (combining predictions for all five cross-validation

replicates) was then used to calculate performance metrics (accuracy, the area under the receiver operating characteristic (ROC) curve, AUC and F1-score for each class) for each type of model. When necessary, these measures were compared using non-parametric tests (Mann-Whitney-Wilcoxon tests in the case of independent samples – between data sets – and Wilcoxon signed-rank tests in the case of comparisons involving the same cross validation replicates, that is, within data sets).

We further explored the possibility of using ensemble models to classify images. To this purpose, we combined the predictions of models for each image by calculating the arithmetic mean of the probability for each class across the three different models. Finally, we went a step further and, for all individuals for which our data set included both dorsal and head lateral images, we combined the estimates in order to produce a more representative prediction. This was done in two different ways: 1) calculating the arithmetic mean of the probability for each class across all six models (three architectures for each of the two views); 2) using only the predictions obtained for the model with the best performance for each perspective and combining the two estimates as above.

Finally, Grad-CAM (Selvaraju et al., 2017, 2020) was used to produce heatmaps showing the areas of each training image that are important in classification. This was performed 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.

3.4.2 Distinction between the nine species in the Iberian and North African clade

The overall workflow was highly similar to the methodology adopted for the two-class problem detailed in section 3.4.1. We used the same cross.validation set up (60% for training + 20% for validation + 20% for testing) and the same three architectures (InceptionV3, ResNet50 and InceptionResNetV2) pre-trained from ImageNet, with the same structure except, of course, for the final layer (9 units, and using the softmax activation function to produce the final predictions). However, there were important differences, of which we highlight:

- 1) the data augmentation protocol was adjusted. Because initial experiments 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 increased to 0.7, and `brightness_range`, which was established between 0.2 and 1.8. This methodology was applied in only three out of the four data sets; in the female dorsal image set this data augmentation set up produced very low accuracies, in the order of 4 to 10% (results not shown; this test was performed only for a few rounds of cross validation for InceptionV3, the fastest-running model). Therefore, in the case of female dorsal images we used exactly the same data augmentation procedure used for the two-class problem. Comparisons involving this model are therefore not completely straightforward because of this difference.
- 2) batch size was increased to 64 to reduce computational time, after verifying that no significant differences in accuracy were observed.
- 3) number of epochs was increased to 2000 because the learning curves took longer to stabilize.

Predictions based on model combinations were performed using the methodology described before for the two-class case (involving both each type of image for each individual and combinations of the two perspectives).

Data evaluation in this case relied in determining accuracy and the F1-score for each class, as well as macro and weighted-averaged over classes. Confusion matrices were plotted in order to visualize classification dynamics.

4. Results

4.1 Distinction between *P. bocagei* and *P. lusitanicus*

The overall performance of the three methods for image classification of *Podarcis bocagei* and *Podarcis lusitanicus* in the four different datasets is shown in table 3. Detailed results including training, validation and test set evaluation for all cross-validation sets are shown in Appendix 2. Accuracy was 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 InceptionResNetV2 on both male dorsal and head lateral images. F1-scores show that typically *P. lusitanicus* was more mis-classified than *P. bocagei*, for both types of images and for both sexes.

All three methods performed similarly in all data sets considering the three-performance metrics.

Identification of males was generally more accurate than that of females. Considering all five cross-validation replicates of the three models, identification accuracy of males was significantly higher than that of females only when considering dorsal images ($p=0.048$, Mann-Whitney-Wilcoxon test). The same result was obtained, but even more pronounced, using other metrics ($p=0.008982$ and $p=0.0298$ for AUC and F1-scores, respectively). With respect to head lateral images the difference between sexes also exists but is significant only for differences in AUC ($p=0.046$, Mann-Whitney-Wilcoxon test).

There was no difference in performance using different perspectives (dorsal or head lateral views), neither in the case of males nor females.

Table 3. Evaluation of the three tested architectures in the four datasets for the two-class problem. Models with the highest accuracy are highlighted in bold.

Sex	View	Metric ^a	Inception V3	ResNet 50	Inception ResNetV2
Males	Dorsal	Accuracy	0.935	0.922	0.948
		AUC	0.976	0.982	0.984
		F1 <i>Pboc</i>	0.951	0.941	0.962
		F1 <i>Plus</i>	0.905	0.887	0.919
	Head lateral	Accuracy	0.926	0.929	0.936
		AUC	0.972	0.975	0.984
		F1 <i>Pboc</i>	0.946	0.947	0.953
		F1 <i>Plus</i>	0.882	0.895	0.889
Females	Dorsal	Accuracy	0.873	0.906	0.905
		AUC	0.931	0.965	0.970
		F1 <i>Pboc</i>	0.905	0.930	0.929
		F1 <i>Plus</i>	0.810	0.857	0.859
	Head lateral	Accuracy	0.935	0.919	0.927
		AUC	0.962	0.959	0.976
		F1 <i>Pboc</i>	0.953	0.941	0.947
		F1 <i>Plus</i>	0.897	0.87	0.872

^a-AUC refers to the area under the ROC curve.

As an extension to this basic approach, we tested whether ensemble models (calculated by averaging predictions of different models) would increase classification success. These results are presented in table 4. Within each of the four data sets, ensemble models do not always improve classification success compared to the best single model. For instance, in the case of head lateral images prediction performance is worse with the ensemble model than when using the best performing model only. In the case of dorsal images, the improvement is very slight for males and more substantial for females.

However, combining the predictions from different views results in a much higher classification success in all cases, particularly when using the combination of all six models available for a particular individual; in this case accuracy reaches as high as 97.1% for males and 95.9% for females.

Table 4. Assessing the utility of ensemble models for image classification in the two-class problem.

Sex	View	Model	Accuracy	AUC	F1 <i>Pboc</i>	F1 <i>Plus</i>
Males	Dorsal	Best single	0.948	0.984	0.962	0.919
		Ensemble	0.955	0.982	0.967	0.930
	Headlat	Best single	0.936	0.984	0.953	0.889
		Ensemble	0.930	0.976	0.949	0.885
	Combined views	Ensemble 6 models	0.971	0.997	0.979	0.953
		Ensemble 2 best	0.961	0.987	0.972	0.937
Females	Dorsal	Best single	0.906	0.965	0.930	0.857
		Ensemble	0.943	0.970	0.958	0.908
	Headlat	Best single	0.935	0.962	0.953	0.897
		Ensemble	0.911	0.972	0.937	0.847
	Combined views	Ensemble 6 models	0.959	0.992	0.970	0.934
		Ensemble 2 best	0.947	0.977	0.962	0.912

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 were indeed considering lizard images for the 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 4. In dorsal images the model often uses the middle area of the trunk (possibly due to the striking patterning) to discriminate the two classes, but the head region was also used (as well as both regions combined). In female dorsal images the head was not as used as the trunk, but the portion of the trunk used for discrimination was generally more anterior than in males. In both male and female head lateral images it was the area around the ear that was mostly used for correct classification, although this region could be more or less shifted towards the throat in both sexes.

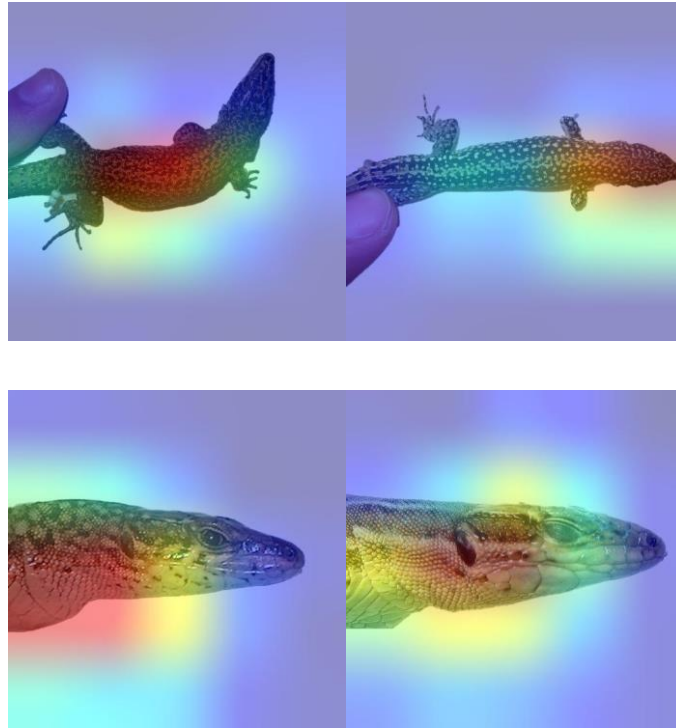


Figure 4. Example of Grad-CAM heatmaps obtained for *Podarvis lusitanicus*. The upper images show two common patterns observed in male dorsal images (also found, albeit with some differences in females). The bottom images exhibit the patterns most frequently found in male and female head lateral images (here illustrated in two females).

4.2 Distinction between all nine species in the Iberian and North African clade

Overall, the performance of the different models for classification over the nine classes was worse than in the two-class case. Unlike runs involving only *P. bocagei* and *P. lusitanicus*, in all cases considering nine classes overfitting was very evident (see Appendix 2 for detailed training, validation and testing evaluation scores). This problem was minimized by experimenting with various options (varying the learning rate and batch size, changing data augmentation procedures, increasing the number of epochs, amongst other experiments) but it could not be completely overcome. Despite training accuracy rapidly arriving to a fixed value around 1, most validation and test set accuracies were well below this value (generally around 80% or even lower in the case of female images). In general, accuracies

ranged from 72.4% for InceptionV3 in female dorsal perspectives up to 85.3% for InceptionResNetV2 in male dorsal views.

A summary of the performance of each model is presented in table 5 and more detailed information for each cross-validation pass are shown in Appendix 2.

Table 5. Evaluation of the three tested architectures in the four datasets for the nine-class problem. Models with the highest accuracy are highlighted in bold.

Sex	View	Metric	Inception V3	ResNet 50	Inception ResNetV2
Males	Dorsal	Accuracy	0.849	0.832	0.853
		F1 macro	0.826	0.806	0.829
	Head lateral	Accuracy	0.832	0.840	0.828
		F1 macro	0.811	0.817	0.807
Females	Dorsal	Accuracy	0.724	0.766	0.738
		F1 macro	0.705	0.753	0.68
	Head lateral	Accuracy	0.783	0.763	0.804
		F1 macro	0.744	0.727	0.775

A striking result was the highly significant difference between male and female image identification accuracy, 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, both for males and females.

In terms of classification ability, models were very similar. There were no major differences between models in classification ability (the only significant difference was detected in female head lateral images, in which ResNet50 performed significantly worse than Inception ResNet V2 ($p = 0.0325$ for both accuracy and F1-score, Wilcoxon signed rank test)).

To study how classification errors are distributed and investigate the contribution of different classes overall, we plotted the confusion matrix for each data set and model, as

well as the distribution of F1-scores across species. These results are presented in Figures 5-8 (for the best model in each case).

These results highlight that several species appear to be fairly well recognisable independently of the data set. *P. carbonelli* is a case in point, particularly for male images. The main problematic species are *P. gadarramae* and *P. liolepis*, and they appear to be the main reason why classification performance is overall only average. *P. gadarramae* is often mistaken for *P. virescens* or, to a lesser extent, *P. lusitanicus* or *P. bocagei*, depending on the image type, whereas individuals from *P. liolepis* are often confused with *P. virescens* or *P. vaucheri* (but also with other species). *P. hispanicus* is also not recovered consistently in some datasets, particularly in male images, with some confusion towards mainly *P. vaucheri*, or to a lesser extent *P. liolepis*, depending on the data set.

An evaluation of ensemble models was also performed in this case. These results are shown in table 6.

Table 6. Assessing the utility of ensemble models for image classification in the nine-class problem

Sex	View	Model	Accuracy	F1 macro
Males	Dorsal	Best single	0.853	0.829
		Ensemble	0.886	0.866
	Headlat	Best single	0.840	0.817
		Ensemble	0.876	0.854
	Combined views	Ensemble 6 models	0.935	0.923
		Ensemble 2 best	0.907	0.882
Females	Dorsal	Best single	0.766	0.753
		Ensemble	0.817	0.790
	Headlat	Best single	0.804	0.775
		Ensemble	0.830	0.802
	Combined views	Ensemble 6 models	0.897	0.880
		Ensemble 2 best	0.866	0.844

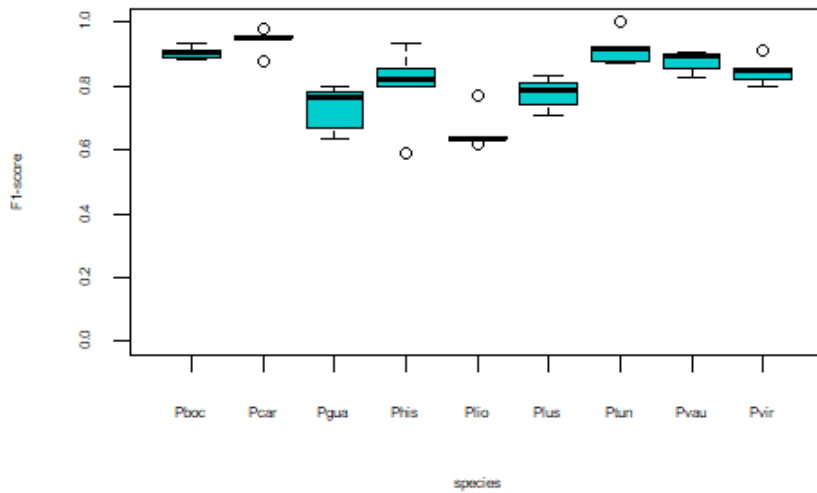
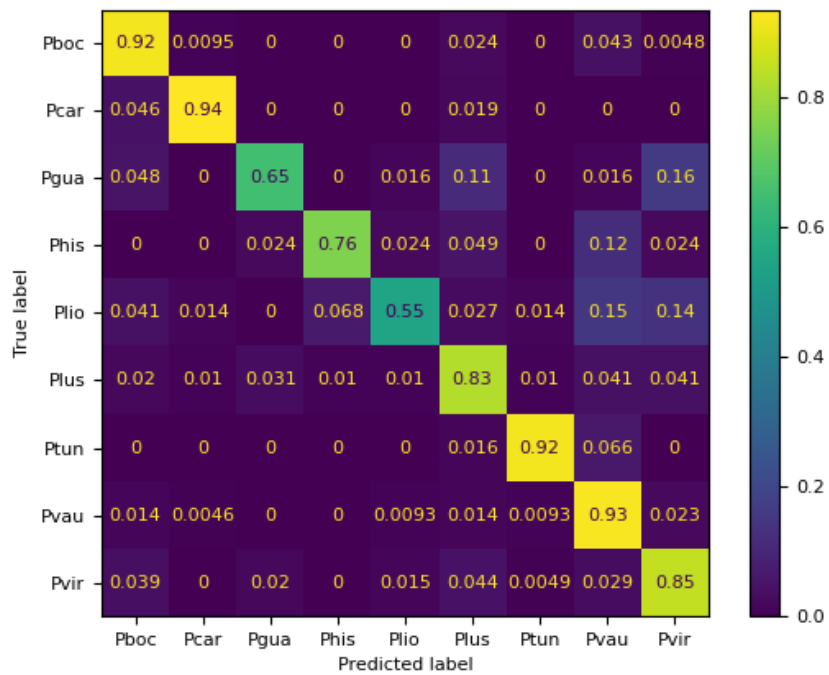


Figure 5. Classification success for male dorsal images based on InceptionResNetV2 results. Upper: confusion matrix normalized over rows; lower: a boxplot describing variation of f1-scores for different classes (based on five cross-validation test sets). Abbreviations used: Pboc, *P. bocagei*; Pcar, *P. carbonelli*; Phis, *P. hispanicus*; Plio, *P. liolepis*; Plus, *P. lusitanicus*; Ptun, *P. tunesiacus*; Pvau, *P. vaucheri*; Pvir, *P. virescens*.

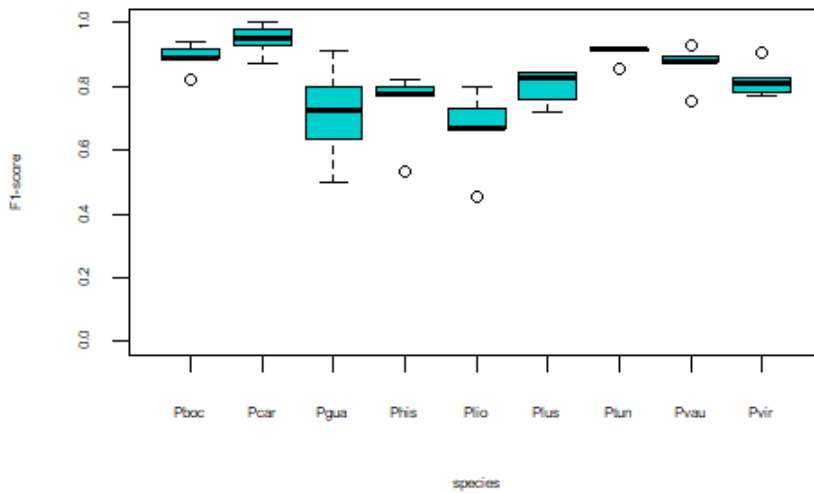
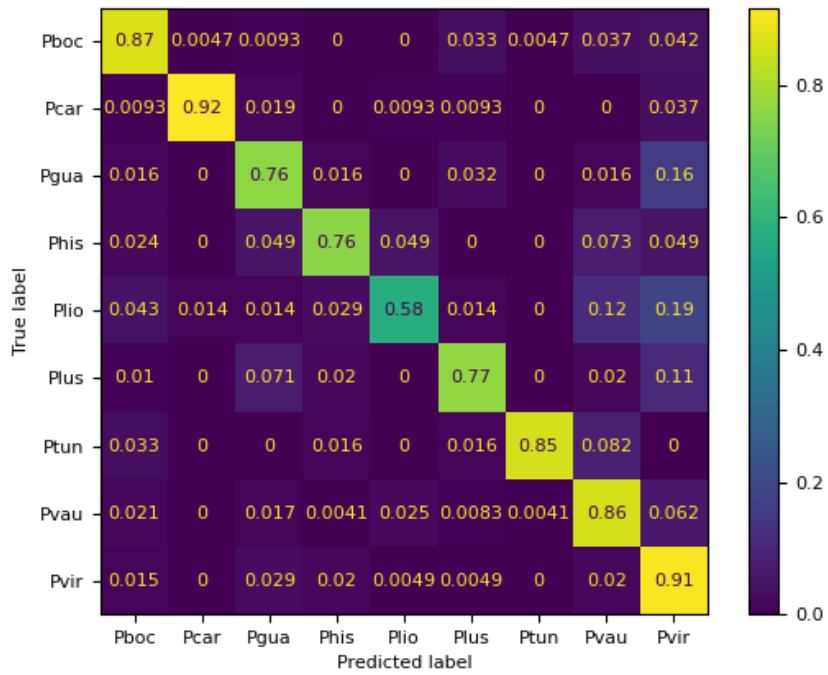


Figure 6. Classification success for male head lateral images based on ResNet50 results. Upper: confusion matrix normalized over rows; lower: a boxplot describing variation of f1-scores for different classes (based on five cross-validation test sets). Abbreviations as in Figure 5.

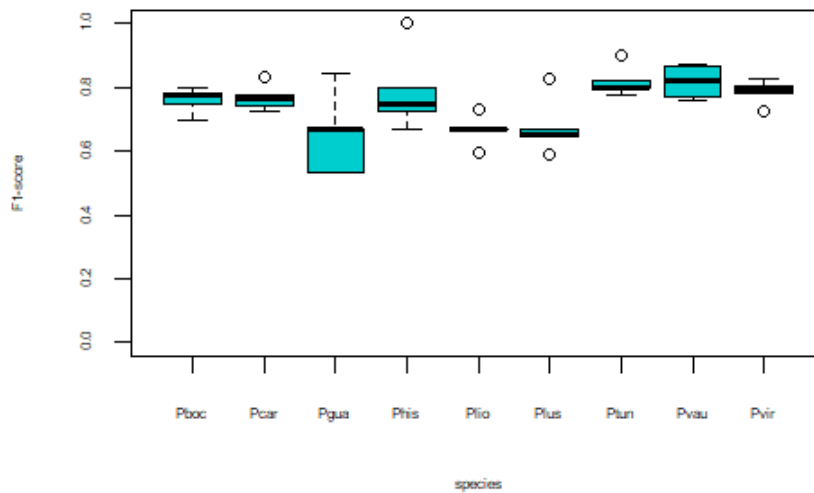
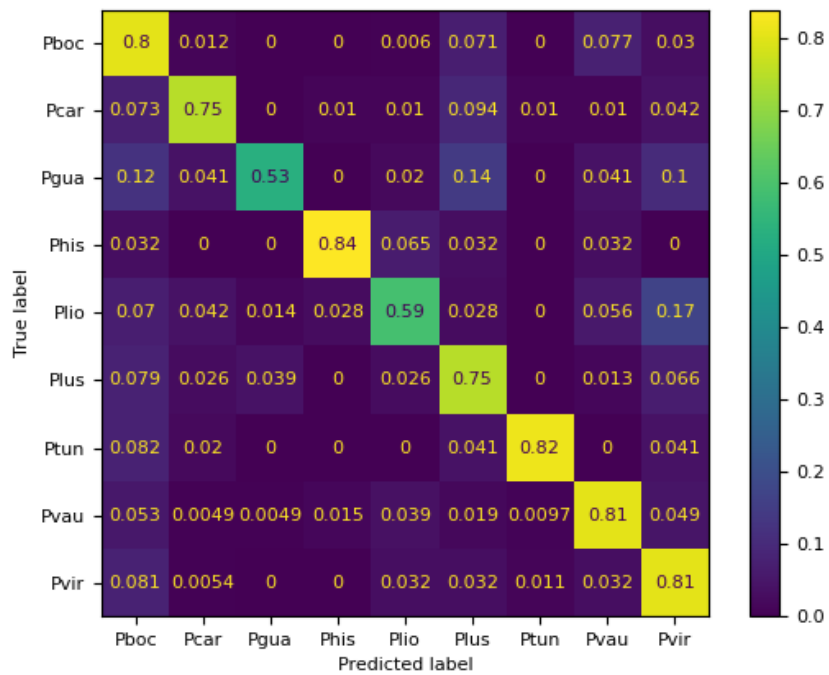


Figure 7. Classification success of female dorsal images based on ResNet50 results. Upper: confusion matrix normalized over rows; lower: a boxplot describing variation of f1-scores for different classes (based on five cross-validation test sets). Abbreviations as in Figure 5.

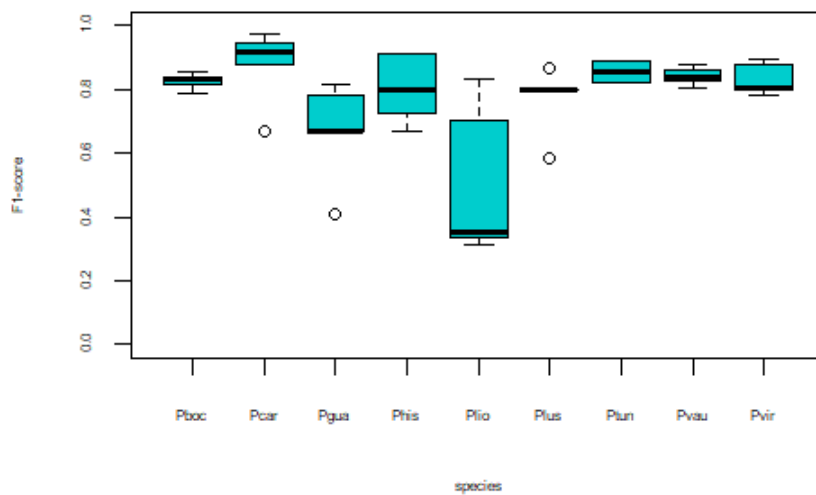
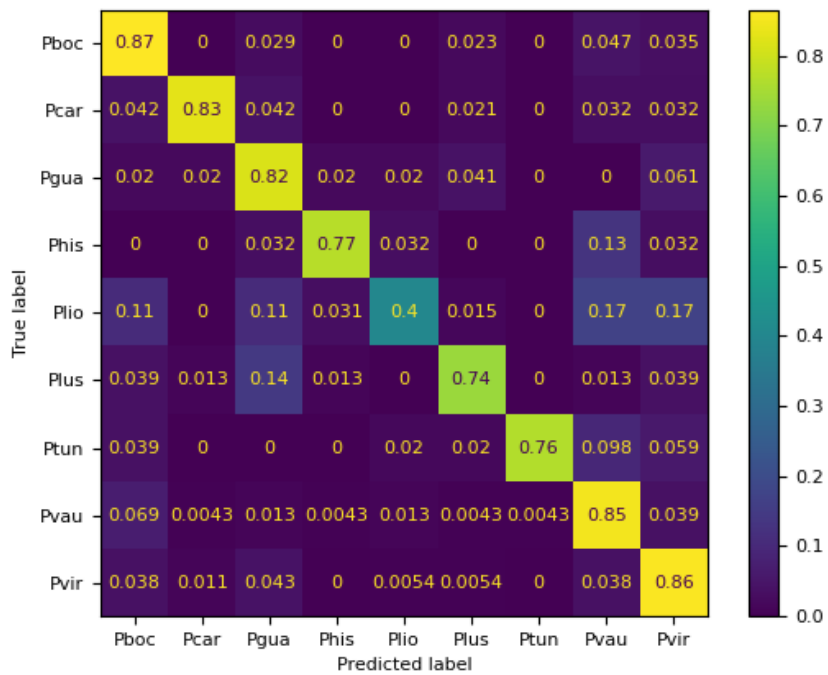


Figure 8. Classification success for female head lateral images based on Inception ResNet V2 results. Upper: confusion matrix normalized over rows; lower: a boxplot describing variation of f1-scores for different classes (based on five cross-validation test sets). Abbreviations as in Figure 5.

Unlike the two-class case, in which the utility of ensemble models was mostly restricted to the combination of predictions from different perspectives, without important improvements in the within-data set case, in the nine-class problem ensemble models greatly improve predictions in all cases, both when combining predictions within data sets (ensemble models always improve estimates from the best single model) but also, and most importantly, when using estimates from different views. In this case, prediction accuracy reaches as high as 93.5% for males and 89.7% for females, both in the case of combining the six models. As in the two-class case, combining only the best model for each perspective increases prediction accuracy but only modestly.

Confusion matrices for the six-model ensemble for males and females are shown in figure 9. These analyses highlight that the classification problems for some species were attenuated by the use of multiple predictions (e.g. for *P. guadarramae*); however, *P. liolepis* is still problematic, both for males and females, with images of this species being erroneously classified as *P. virescens* or *P. vaucheri*. The classification of *P. hispanicus* and *P. lusitanicus* improved only slightly in comparison to others.

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 evaluate what the model considers for discrimination. More precisely, the same regions seem to be used to classify distinct species, but it is not quite evident how differences in these regions are used. The most common patterns found for each species are summarized in Tables 7 and 8 (for males and females, respectively). Curiously, the tip of the snout is frequently used to classify female images, whereas this region is typically irrelevant in male images.

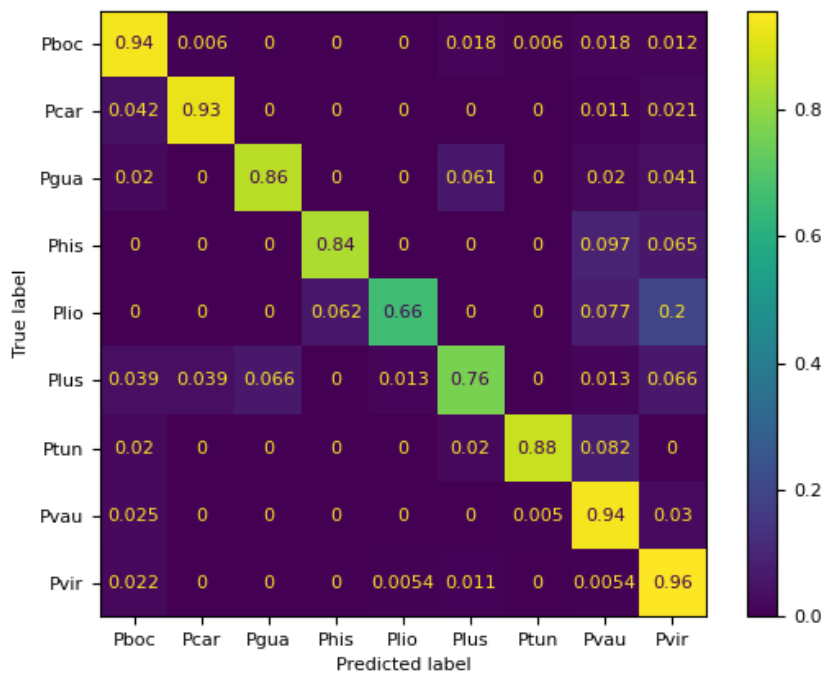
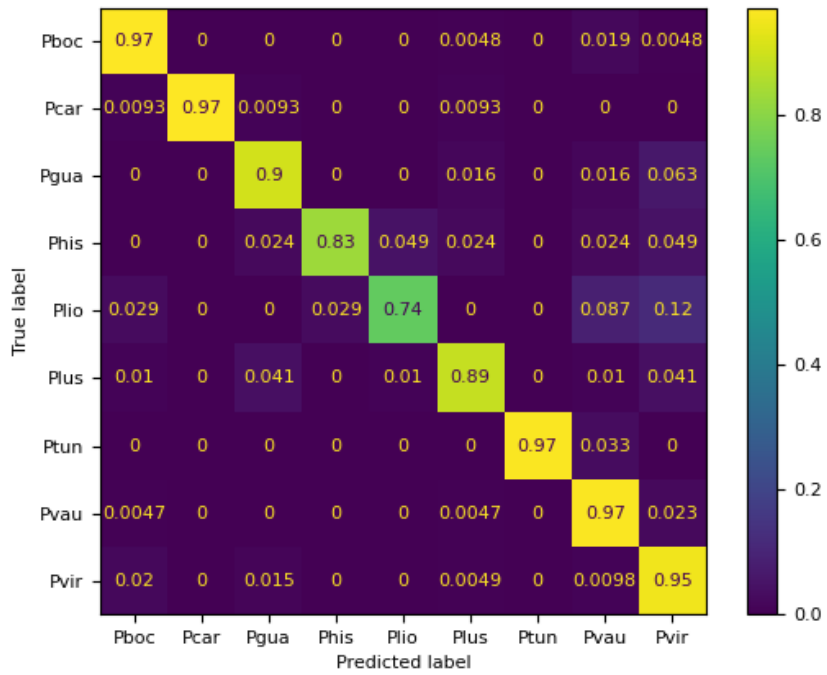


Figure 9. Confusion matrix for male (upper) and female (lower) image classification based on a combination of predictions from the six models applied. Abbreviations as in Figure 5.

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

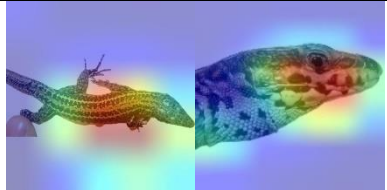
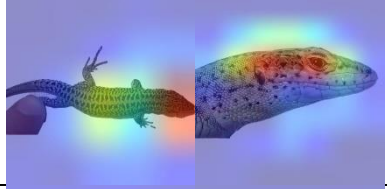

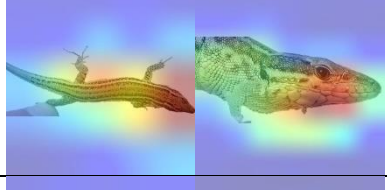

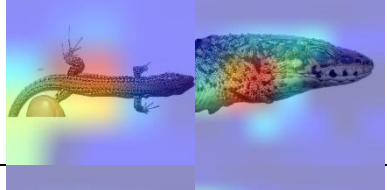
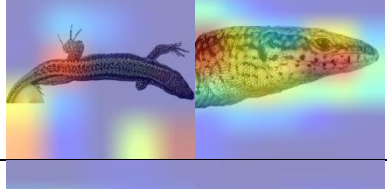
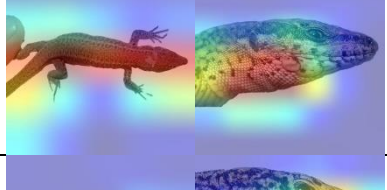


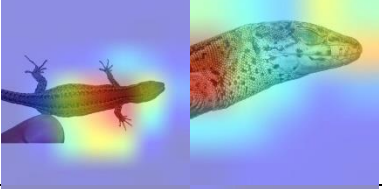

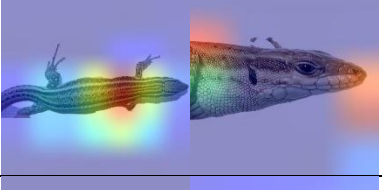

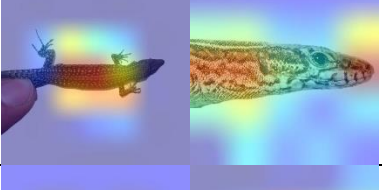
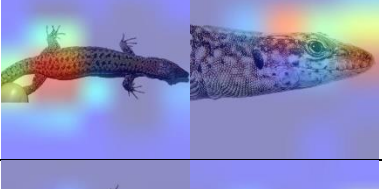
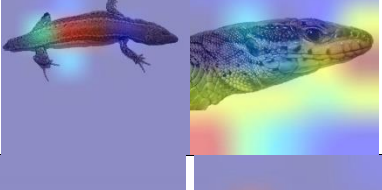
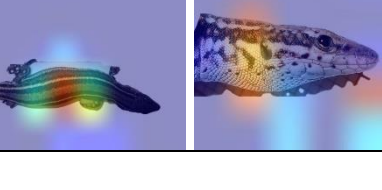
<i>P. bocagei</i>	Highly variable (no clear pattern). All portions of the dorsal views were equally used. In head images the area around the eye, the top of the head, the snout and the throat were all used in similar proportions.	
<i>P. carbonelli</i>	Variable for both views. Snout and middle of the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

Table 8. Summary of Grad-CAM results for each class (females)

<i>P. bocagei</i>	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.	
<i>P. carbonelli</i>	Variable. In the dorsal view, the region around the anterior limb insertion is frequently used. In the head lateral view, the tip of the snout is commonly used, as well as the most posterior region of the head.	
<i>P. guadarramae</i>	Variable. Mid portion of the dorsum and tip of the snout are the regions used more frequently in dorsal and head lateral views, respectively.	
<i>P. hispanicus</i>	Variable. Different regions of the dorsum, but all where the striped pattern is obvious, are used. Snout and/or top of posterior region of head used.	
<i>P. liolepis</i>	Variable. Anterior part of the dorsum more used than other regions, whereas the tip of the snout is used in most head lateral images.	
<i>P. lusitanicus</i>	Mid dorsum, in the dorsal view, and both snout and posterior side of the head (in head lateral views) frequently used.	
<i>P. tunesiacus</i>	Variable. Posterior part of the dorsum more used than in other species; snout and top head region behind the eye used with some frequency.	
<i>P. vaucheri</i>	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.	
<i>P. virescens</i>	Highly variable. All portions of the dorsum used (but in generally small areas) in dorsal images, region around and behind the ear more used than in other species for head lateral images.	

5. Discussion

In this study, we proposed the application of deep learning algorithms to the identification of images belonging to closely related and morphologically similar lizard species. Taxonomically speaking, to our knowledge this is one of the first studies of the kind conducted in squamates, and the first not involving snakes. The objects of this study, wall lizards belonging to the *P. hispanicus* species complex, are common, widespread species frequently found across Iberia and the Maghreb, and have a long tradition of being a challenge for taxonomists because of the combination of low interspecific with huge inter-individual morphological variation (Kaliontzopoulou, et al., 2012b). Studies focusing on automating the identification of other biological species face different challenges (e.g. the very high number of classes or of images), but to our knowledge no study had yet focused on distinguishing images of species that are so similar morphologically as the objects of this study.

We conducted this study in a two-stage process: first we addressed a simple problem, both in the number of classes (only two) and in the degree of morphological differentiation (relatively high), and then moved on to a more complex problem involving nine classes including completely cryptic forms (that is, species that were only described based on genetics due to the impossibility of morphological distinction even by experts).

*High classification success in the discrimination between *P. bocagei* and *P. lusitanicus**

With respect to the first problem, the distinction between *P. bocagei* and *P. lusitanicus*, the performance of computer vision models was high. This was expected, since these two species show morphological differences that enable their distinction by experts (namely a smaller size, less intense green in the dorsum and flatter heads in the case of *P. lusitanicus*); however, compared to other species that have been the object of studies involving deep learning tools, they can still be considered fairly cryptic. 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 combining results from the two perspectives) is comparable to the accuracies generally reported in similar studies (see Table 1). It should be noted that we report testing accuracies and a large portion of the studies report validation accuracy, based on which the models are optimized during learning.

The complex nine-class scenario

When we moved to a more complex problem, the distinction of the nine species in the Iberian and North African clade, classification success dropped significantly. For single models, accuracy ranged from 76.6% to 85.3% for the best performing models applied to female dorsal and male dorsal images, respectively. Combining the predictions for each view increased this success to a moderate extent, and classification was highest when combining predictions from different views for the same individual using all six models combined (93.5% in the case of males and 89.7% 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*.

This result is largely led 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 very successfully classified (particularly, but not only, in ensemble models), but individuals from other species are also frequently mistaken, and this is not completely solved by combining predictions from different models and views. *P. liolepis*, the Catalanian wall lizard, is a case in point. This is a species with a vast distribution throughout the eastern half of the Iberian Peninsula, likely one of the most widespread among those included in this study (see distribution map in <https://www.eurolizards.com/lizards/podarcis-liolepis/>). This probably means that this species encompasses a great deal of morphological variability resulting from adaptation or developmental plasticity to cope with widely different climatic conditions. Moreover, some southern populations are completely isolated from the remainder of the species, suggesting that genetic drift might accentuate differentiation among populations. Although overall morphological patterns might be diverse, *P. liolepis* individuals have simpler dorsal patterns compared to those that other species exhibit (A. Kaliontzopoulou, pers. comm). If these patterns are used by models to discriminate species (as it appears from Grad-CAM results), it may happen that this lack of complexity hampers

the identification of this species. Finally, unlike other lizard species, which tend to be more or less homogeneous genetically, *P. liolepis* includes two very 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. Coupled with the relatively low sample size available for this species (which nevertheless includes individuals from different, geographically widespread locations), these factors might result in a particularly difficult problem to be tackled by classification models. If this hypothesis is correct, increasing the sample size for this species in order to become more representative of intraspecific morphological variation could have a positive effect on classification outcomes. An also important probable reason for the misclassification of *P. liolepis* and other species is also current gene flow. This is a feature very common in all of genus *Podarcis* (Yang et al., 2021), and the *P. hispanicus* complex is no exception (see e.g. Caeiro-Dias et al., 2020). This phenomenon could have a strong impact on the morphology of some individuals, particularly those coming from regions near contact zones.

A quite unexpected result of this study is the relatively high ability (considering the prior expectations) for the models to distinguish between *P. lusitanicus* and *P. gadarramae*. This is the only truly 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). As such, they were first described as subspecies and only recently elevated to the species status, after studies on their genetic variation clarified their distinctiveness (G. Caeiro-Dias et al., 2021). It is thus remarkable that in our study the proportion of individuals of one species identified as the other is as low as 1.6% - 6.7% using ensemble models. An interesting follow-up study will thus be an analysis of classification focusing only on these two species.

An also interesting 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 that is currently of conservation concern (having been classified as “endangered” by the IUCN). 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 programs directed at this species once these models become available in naturalist mobile applications.

It should be emphasised 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 thesis are still, by far, the 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 lacertid species: biometry (linear body measurements), pholidotic (scale-count) characters and a combination of both types of characters. Although some of the classes considered in this work were now merged in our study (since the knowledge about species limits has improved), and hence the results cannot be straightforwardly compared, classification results for each class 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 could also be a future possibility to consider improving even further the use of computer vision models.

Sexual dimorphism and classification success

Lizards of genus *Podarcis* typically exhibit a marked sexual dimorphism; that is, males and females show strikingly different morphology. Sexual dimorphism in this group translates into differences being more pronounced among males of different species than between females of different species. Females are usually more uniform since they are less brightly colored and lack other external features that typically help in the identification of males (Kaliontzopoulou et al., 2007). In similarity to the difficulties experienced by human observers, our results highlight that classification success was lower in females than in males in both problems. In the two-class problem, classification success differences were only evident in the dorsal view, whereas in the nine-class problem these differences were very significant for both views. The fact that sexual dimorphism does not affect the distinction between head lateral images of *P. bocagei* and *P. lusitanicus* females likely results from the fact that the most obvious difference among the two species, the high degree of head flattening exhibited by *P. lusitanicus* likely related to adaptation to living in rock crevices (Gomes et al., 2016; Kaliontzopoulou, et al., 2012a; Kaliontzopoulou et al., 2012b),

is shared by both males and females (Grad-CAM results suggest that the height of the head is indeed a feature the models consider). However, this feature alone does not work in distinguishing 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, but still higher than that of models applied to dorsal images, where females are much more similar among species.

Curiously, overall patterns of correct classification appear to be common between the two sexes, e.g. *P. liolepis* is poorly classified both in the case of males and females. This suggests that at least some of the characters that the models are looking at characters are not sexually dimorphic.

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

The visualization of heatmaps produced by Grad-CAM allowed highlighting regions that were used by the algorithms for deciding between the classes. On one hand, this was important to verify that the models were looking at lizards and not irrelevant aspects of the images. On the other hand, the goal of attempting to decipher the features in lizard images that differ the most between species (and that could constitute valuable new knowledge in terms of understanding the eco-evolutionary 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 likely different aspects of these regions were used for classification but not providing hints on which particular aspects these were. 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. An intriguing feature is the nearly complete absence of the use of the tip of the snout to classify male images, while this is a recurrent feature in female heatmaps. Because Grad-CAM analysis was only performed for a single model in each case, it remains to be evaluated whereas this is a feature related to the particular model that was chosen to produce heatmaps (the best in each case) or if it is indeed a feature related to general female image classification (and hence of real biological significance).

Methodological considerations

Despite our best efforts, classification success for the second problem was only moderate. 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. 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.

An interesting observation deriving from this study is that there are no major differences in success when applying different deep learning models. The three models used in this work differ in the depth and in the general architecture of the convolutional neural networks. Despite these differences they all perform rather similarly on the data sets. 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 is possible that a different type of ensemble model framework (e.g. combining features retrieved from different models and using them in a machine-learning context) could result in an even bigger improvement. Nevertheless, running multiple models involves a considerable computational cost and may not be feasible for general analyses in the long term.

Finally, a special consideration involves the possible use of different perspectives in the same framework, which predictably will result in much larger classification successes. Marques et al., (2018) suggest a wide array of ensemble approaches to tackle this problem which can be a starting point, but even combining the two different perspectives into a single image might have produced important improvements. Due to time limitations, these improvements were not fully explored in this thesis, but they could certainly be the object of future work.

6. Conclusions and future perspectives

With this work we have shown that deep learning models can be successfully used for the identification of wall lizard species, achieving classification accuracies likely comparable to that of experienced observers and larger than of the common citizen (including that of the author of this thesis). However, in the more complex scenario explored in this thesis (a nine-class classification problem) the error rate is still moderate, and misclassification instances are still high for some species such as *P. liolepis*; therefore, the practical deployment of these models for research or conservation purposes is still not a possibility. However, this work suggests that there are prospects for improving the success of these models, such as augmenting class sample sizes to achieve a balanced data set and extending the use of ensemble models.

Another practical consideration to include in future developments will be the use of geographical information; in this work we took the challenge of discriminating between all nine classes simultaneously for academic purposes, but this is a problem that a naturalist will not face in the field; although there are species that overlap and regions where the distribution is not well-known, a real-life problem will involve distinguishing between at most 3-4 species simultaneously. Including geographical coordinate information will thus certainly facilitate the classification problem.

In general, beyond the specific problem of classifying wall lizards, this work shows that computer vision models can work 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 (that nevertheless remain to be evaluated).

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. *12th symposium on operating systems design and implementation (16)*, 265–283.
- Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.-C., & Joly, A. (2017). *Pl@ntnet app in the era of deep learning*.
- Almryad, A. S., & Kutucu, H. (2020). Automatic identification for field butterflies by convolutional neural networks. *Engineering Science and Technology, an International Journal*, 23(1), 189–195.
- Arzar, N. N. K., Sabri, N., Johari, N. F. M., Shari, A. A., Noordin, M. R. M., & Ibrahim, S. (2019). Butterfly Species Identification Using Convolutional Neural Network (CNN). *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 221–224.
- Baldi, P., & Sadowski, P. J. (2013). Understanding dropout. *Advances in neural information processing systems*, 26, 2814–2822.
- Banan, A., Nasiri, A., & Taheri-Garavand, A. (2020). Deep learning-based appearance features extraction for automated carp species identification. *Aquacultural Engineering*, 89, 102053.
- Barré, P., Stöver, B. C., Müller, K. F., & Steinhage, V. (2017). LeafNet: A computer vision system for automatic plant species identification. *Ecological Informatics*, 40, 50–56.
- Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K., Meier, R., Winker, K., Ingram, K. K., & Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in ecology & evolution*, 22(3), 148–155.
- Bonnet, P., Joly, A., Faton, J.-M., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A., Lombardo, J.-C., & Mary, L. (2020). How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. *Ecological Solutions and Evidence*, 1(2), e12023.
- Buhrmester, V., Münch, D., & Arens, M. (2019). Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv preprint arXiv:1911.12116*.
- 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., Brelsford, A., Kaliontzopoulou, A., Meneses-Ribeiro, M., Crochet, P.-A., & Pinho, C. (2020). Variable levels of introgression between the endangered *Podarcis carbonelli* and highly divergent congeneric species. *Heredity*, 1–14.
- Caeiro-Dias, G., Luís, C., Pinho, C., Crochet, P.-A., Sillero, N., & Kaliontzopoulou, A. (2018). *Lack of congruence of genetic and niche divergence in Podarcis hispanicus complex*. Wiley Online Library.

- Caeiro-Dias, G. M. C. (2018). *Understanding speciation: A multidisciplinary assessment of hybrid zones using a lizard species complex as model*.
- Caeiro-Dias, G., Rocha, S., Couto, A., Pereira, C., Brelford, A., Crochet, P.-A., & Pinho, C. (2021). Nuclear phylogenies and genomics of a contact zone establish the species rank of *Podarcis lusitanicus* (Squamata, Lacertidae). *Molecular Phylogenetics and Evolution*, *164*, 107270. <https://doi.org/10.1016/j.ympev.2021.107270>
- Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science advances*, *1*(5), e1400253.
- Chen, G., Han, T. X., 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)*, 858–862.
- Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv:1610.02357 [cs]*. <http://arxiv.org/abs/1610.02357>
- Chollet, F. & others. (2018). Keras: The python deep learning library. *Astrophysics Source Code Library*, ascl-1806.
- Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. *Twenty-second international joint conference on artificial intelligence*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.
- dos Santos, A. A., & Goncalves, W. N. (2019). Improving Pantanal fish species recognition through taxonomic ranks in convolutional neural networks. *Ecological Informatics*, *53*, 100977.
- dos Santos, A. A., & Gonçalves, W. N. (2019). Improving Pantanal fish species recognition through taxonomic ranks in convolutional neural networks. *Ecological Informatics*, *53*, 100977.
- Drew, L. W. (2011). Are We Losing the Science of Taxonomy?: As need grows, numbers and training are failing to keep up. *BioScience*, *61*(12), 942–946. <https://doi.org/10.1525/bio.2011.61.12.4>
- Gaston, K. J., & O'Neill, M. A. (2004). Automated species identification: Why not? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 655–667.
- Geniez, P., Cluchier, A., Sá-Sousa, P., Guillaume, C. P., & Crochet, P.-A. (2007). Systematics of the *Podarcis hispanicus*-complex (Sauria, Lacertidae) I: Redefinition, morphology and distribution of the nominotypical taxon. *The Herpetological Journal*, *17*(2), 69–80.

- Geniez, P., Sa-Sousa, P., Guillaume, C. P., Cluchier, A., & Crochet, P.-A. (2014). Systematics of the *Podarcis hispanicus* complex (Sauria, Lacertidae) III: Valid nomina of the western and central Iberian forms. *Zootaxa*, 3794(1), 1–51.
- Ghiselin, M. T. (2001). Species concepts. *e LS*.
- Gogul, I., & Kumar, V. S. (2017). Flower species recognition system using convolution neural networks and transfer learning. *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*, 1–6.
- Gomes, V., Carretero, M. A., & Kaliontzopoulou, A. (2016). The relevance of morphology for habitat use and locomotion in two species of wall lizards. *Acta Oecologica*, 70, 87–95. <https://doi.org/10.1016/j.actao.2015.12.005>
- Gómez-Ríos, A., Tabik, S., Luengo, J., Shihavuddin, A. S. M., & Herrera, F. (2019). Coral species identification with texture or structure images using a two-level classifier based on Convolutional Neural Networks. *Knowledge-Based Systems*, 184, 104891. <https://doi.org/10.1016/j.knosys.2019.104891>
- Gómez-Ríos, A., Tabik, S., Luengo, J., Shihavuddin, A. S. M., Krawczyk, B., & Herrera, F. (2019). Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation. *Expert Systems with Applications*, 118, 315–328.
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.
- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187, 27–48.
- Hansen, O. L., Svenning, J.-C., Olsen, K., Dupont, S., Garner, B. H., Iosifidis, A., Price, B. W., & Høye, T. T. (2020). Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and evolution*, 10(2), 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.
- Hey, J. (2001). *Genes, categories, and species: The evolutionary and cognitive cause of the species problem*. Oxford University Press.
- Hey, J., Waples, R. S., Arnold, M. L., Butlin, R. K., & Harrison, R. G. (2003). Understanding and confronting species uncertainty in biology and conservation. *Trends in Ecology & Evolution*, 18(11), 597–603.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580 [cs]*. <http://arxiv.org/abs/1207.0580>
- Hopkins, G. W., & Freckleton, R. P. (2002). Declines in the numbers of amateur and professional taxonomists: Implications for conservation. *Animal Conservation*, 5(3), 245–249.

- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Hsiang, A. Y., Brombacher, A., Rillo, M. C., Mleneck-Vautravers, M. J., Conn, S., Lordsmith, S., Jentzen, A., Henehan, M. J., Metcalfe, B., & Fenton, I. S. (2019). Endless Forams:> 34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. *Paleoceanography and Paleoclimatology*, *34*(7), 1157–1177.
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*. <http://arxiv.org/abs/1608.06993>
- Kaliontzopoulou, A. (2010). *Proximate and evolutionary causes of phenotypic diversification: Morphological variation in iberian and north african podarcis wall lizards* [PhD Thesis]. Universitat de Barcelona.
- Kaliontzopoulou, A., Adams, D. C., van der Meijden, A., Perera, A., & Carretero, M. A. (2012a). Relationships between head morphology, bite performance and ecology in two species of Podarcis wall lizards. *Evolutionary Ecology*, *26*(4), 825–845.
- Kaliontzopoulou, A., Carretero, M. A., & Llorente, G. A. (2007). Multivariate and geometric morphometrics in the analysis of sexual dimorphism variation in Podarcis lizards. *Journal of Morphology*, *268*(2), 152–165.
- Kaliontzopoulou, A., Carretero, M. A., & Llorente, G. A. (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*(1), 173–193. <https://doi.org/10.1111/j.1096-3642.2011.00760.x>
- Kaliontzopoulou, A., Pinho, C., Harris, D. J., & Carretero, M. A. (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*(4), 779–800.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.
- Lee, S. H., Chan, C. S., Wilkin, P., & Remagnino, P. (2015). Deep-plant: Plant identification with convolutional neural networks. *2015 IEEE international conference on image processing (ICIP)*, 452–456.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11–26. <https://doi.org/10.1016/j.neucom.2016.12.038>
- Lu, Y.-C., Tung, C., & Kuo, Y.-F. (2020a). Identifying the species of harvested tuna and billfish using deep convolutional neural networks. *ICES Journal of Marine Science*, *77*(4), 1318–1329.

- Lu, Y.-C., Tung, C., & Kuo, Y.-F. (2020b). Identifying the species of harvested tuna and billfish using deep convolutional neural networks. *ICES Journal of Marine Science*, 77(4), 1318–1329.
- MacLeod, N., Benfield, M., & Culverhouse, P. (2010). Time to automate identification. *Nature*, 467(7312), 154–155.
- Mäder, P., Boho, D., Rzanny, M., Seeland, M., Wittich, H. C., Deggelmann, A., & Wäldchen, J. (2021). The flora incognita app—interactive plant species identification. *Methods in Ecology and Evolution*.
- Marques, A. C. R., M. Raimundo, M., B. Cavalheiro, E. M., FP Salles, L., Lyra, C., & J. Von Zuben, F. (2018a). Ant genera identification using an ensemble of convolutional neural networks. *Plos one*, 13(1), e0192011.
- Marques, A. C. R., M. Raimundo, M., B. Cavalheiro, E. M., FP Salles, L., Lyra, C., & J. Von Zuben, F. (2018b). Ant genera identification using an ensemble of convolutional neural networks. *Plos one*, 13(1), e0192011.
- Miao, Z., Gaynor, K. M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M. S., McInturff, A., Bowie, R. C., Nathan, R., & Stella, X. Y. (2019). Insights and approaches using deep learning to classify wildlife. *Scientific reports*, 9(1), 1–9.
- Miao, Z., Gaynor, K. M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M. S., McInturff, A., Bowie, R. C., Nathan, R., Stella, X. Y., & others. (2019). Insights and approaches using deep learning to classify wildlife. *Scientific reports*, 9(1), 1–9.
- Miele, V., Dussert, G., Cucchi, T., & Renaud, S. (2020). Deep learning for species identification of modern and fossil rodent molars. *bioRxiv*.
- Milošević, D., Milosavljević, A., Predić, B., Medeiros, A. S., Savić-Zdravković, D., Stojković Piperac, M., 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. <https://doi.org/10.1016/j.scitotenv.2019.135160>
- Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Icml*.
- Nguyen, H., Maclagan, S. J., Nguyen, T. D., Nguyen, T., Flemons, P., Andrews, K., Ritchie, E. G., & 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)*, 40–49.
- Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), E5716–E5725.
- Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in zoology*, 7(1), 1–14.

- Picek, L., Bolon, I., Durso, A. M., & de Castañeda, R. R. (2020). *Overview of the SnakeCLEF 2020: Automatic Snake Species Identification Challenge*. 15.
- Raphael, A., Dubinsky, Z., Iluz, D., Benichou, J. I., & Netanyahu, N. S. (2020). Deep neural network recognition of shallow water corals in the Gulf of Eilat (Aqaba). *Scientific reports*, *10*(1), 1–11.
- Rauf, H. T., Lali, M. I. U., Zahoor, S., Shah, S. Z. H., Rehman, A. U., & Bukhari, S. A. C. (2019). Visual features based automated identification of fish species using deep convolutional neural networks. *Computers and Electronics in Agriculture*, *167*, 105075.
- Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*(9), 2352–2449. https://doi.org/10.1162/neco_a_00990
- Renoult, J. P., Geniez, P., Bacquet, P., Benoit, L., & CROCHET, P.-A. (2009). Morphology and nuclear markers reveal extensive mitochondrial introgressions in the Iberian Wall Lizard species complex. *Molecular ecology*, *18*(20), 4298–4315.
- Renoult, J. P., Geniez, P., Bacquet, P., Guillaume, C. P., & Crochet, P.-A. (2010). Systematics of the *Podarcis hispanicus*-complex (Sauria, Lacertidae) II: The valid name of the north-eastern Spanish form. *Zootaxa*, *2500*(1), 58–68.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.
- Salvi, D., Pinho, C., Mendes, J., & Harris, D. J. (2021). Fossil-calibrated time tree of *Podarcis* wall lizards provides limited support for biogeographic calibration models. *Molecular Phylogenetics and Evolution*, *161*, 107169.
- Seeland, M., Rzanny, M., Boho, D., Wäldchen, J., & Mäder, P. (2019a). Image-based classification of plant genus and family for trained and untrained plant species. *BMC bioinformatics*, *20*(1), 1–13.
- Seeland, M., Rzanny, M., Boho, D., Wäldchen, J., & Mäder, P. (2019b). Image-based classification of plant genus and family for trained and untrained plant species. *BMC Bioinformatics*, *20*(1), 4. <https://doi.org/10.1186/s12859-018-2474-x>
- Sejnowski, T. J. (2018). *The Deep Learning Revolution*. MIT Press.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. <https://doi.org/10.1007/s11263-019-01228-7>

- Selvaraju, R. R., 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.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- 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*, 31(1).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv:1409.4842 [cs]*. <http://arxiv.org/abs/1409.4842>
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]*. <http://arxiv.org/abs/1512.00567>
- The ImageMagick Development Team. (2021). *ImageMagick* (7.0.10) [Computer software]. <https://imagemagick.org>
- Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The iNaturalist Species Classification and Detection Dataset. *arXiv:1707.06642 [cs]*. <http://arxiv.org/abs/1707.06642>
- Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11), 2216–2225.
- Wäldchen, J., Rzanny, M., Seeland, M., & Mäder, P. (2018). Automated plant species identification—Trends and future directions. *PLoS computational biology*, 14(4), e1005993.
- Wilson, E. O. (2004). Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444), 739–739. <https://doi.org/10.1098/rstb.2003.1440>
- Yang, W., Feiner, N., Pinho, C., While, G. M., Kaliontzopoulou, A., Harris, D. J., Salvi, D., & Uller, T. (2021). Extensive introgression and mosaic genomes of Mediterranean endemic lizards. *Nature communications*, 12(1), 1–8.
- Zachos, F. E. (2016). *Species concepts in biology* (Vol. 801). Springer.
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision*, 818–833.
- Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P., Senior, A., Vanhoucke, V., & Dean, J. (2013). On rectified linear units for speech processing. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3517–3521.
- 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)*, 2, 103–106.

Appendix 1. Detailed evaluation results

A1.1 Two-classes case

In all tables below AUC stands for area under the ROC curve and the f1-score refers to that of *P. lusitanicus*.

Table A1. Detailed classification results for male dorsal images

model	cv	training accuracy	training auc	training f1	validation accuracy	validation auc	validation f1	test accuracy	test auc	test f1
incv3	1	1.000	1.000	1.000	0.903	0.948	0.857	1.000	1.000	1.000
incv3	2	0.968	0.995	0.952	0.902	0.951	0.864	0.871	0.951	0.833
incv3	3	1.000	1.000	1.000	0.935	0.994	0.905	0.885	0.966	0.837
incv3	4	1.000	1.000	1.000	0.968	0.960	0.950	0.968	0.998	0.950
incv3	5	0.995	1.000	0.992	0.951	0.979	0.919	0.952	0.993	0.927
mean		0.992	0.999	0.989	0.932	0.966	0.899	0.935	0.981	0.909
stdev		0.014	0.002	0.021	0.029	0.020	0.039	0.055	0.022	0.073
resnet50	1	1.000	1.000	1.000	0.919	0.962	0.872	0.984	0.999	0.974
resnet50	2	1.000	1.000	1.000	0.951	0.977	0.923	0.935	0.975	0.900
resnet50	3	1.000	1.000	1.000	0.903	0.989	0.870	0.934	0.983	0.905
resnet50	4	0.929	1.000	0.899	0.839	0.963	0.792	0.806	0.989	0.769
resnet50	5	1.000	1.000	1.000	0.918	0.997	0.884	0.952	0.996	0.927
mean		0.986	1.000	0.980	0.906	0.978	0.868	0.922	0.989	0.895
stdev		0.032	0.000	0.045	0.041	0.016	0.048	0.068	0.010	0.076
incrv2	1	1.000	1.000	1.000	0.903	0.985	0.833	1.000	1.000	1.000
incrv2	2	1.000	1.000	1.000	0.918	0.984	0.872	0.935	0.975	0.900
incrv2	3	0.995	1.000	0.992	0.903	0.970	0.857	0.918	0.972	0.872
incrv2	4	1.000	1.000	1.000	0.935	0.996	0.905	0.935	0.996	0.909
incrv2	5	1.000	1.000	1.000	0.934	0.994	0.889	0.952	0.998	0.919
mean		0.999	1.000	0.998	0.919	0.986	0.871	0.948	0.988	0.920
stdev		0.002	0.000	0.004	0.016	0.011	0.028	0.031	0.014	0.048

Table A2. Detailed classification results for male head lateral images

model	cv	training accuracy	training auc	training fl	validation accuracy	validation auc	validation fl	test accuracy	test auc	test fl
incv3	1	1.000	1.000	1.000	0.905	0.983	0.850	0.902	0.959	0.824
incv3	2	1.000	1.000	1.000	0.968	0.995	0.950	0.937	0.985	0.905
incv3	3	1.000	1.000	1.000	0.889	0.957	0.837	0.952	0.998	0.930
incv3	4	1.000	1.000	1.000	0.921	0.990	0.872	0.937	0.990	0.905
incv3	5	0.989	1.000	0.983	0.869	0.986	0.733	0.905	0.976	0.824
mean		0.998	1.000	0.997	0.910	0.982	0.848	0.926	0.981	0.877
stdev		0.005	0.000	0.008	0.038	0.015	0.078	0.022	0.015	0.050
resnet50	1	1.000	1.000	1.000	0.937	0.988	0.900	0.918	0.964	0.857
resnet50	2	1.000	1.000	1.000	0.921	0.985	0.889	0.952	0.990	0.930
resnet50	3	1.000	1.000	1.000	0.937	0.993	0.905	0.905	0.999	0.870
resnet50	4	0.963	0.999	0.944	0.937	0.957	0.905	0.905	0.985	0.870
resnet50	5	1.000	1.000	1.000	0.934	0.960	0.889	0.968	0.984	0.950
mean		0.993	1.000	0.989	0.933	0.977	0.897	0.930	0.984	0.895
stdev		0.017	0.000	0.025	0.007	0.017	0.008	0.029	0.013	0.042
incrnv2	1	1.000	1.000	1.000	0.937	0.992	0.895	0.918	0.981	0.857
incrnv2	2	1.000	1.000	1.000	0.952	0.997	0.930	0.921	0.990	0.884
incrnv2	3	1.000	1.000	1.000	0.952	0.991	0.927	0.937	0.988	0.905
incrnv2	4	1.000	1.000	1.000	0.937	0.977	0.905	0.968	0.987	0.947
incrnv2	5	1.000	1.000	1.000	0.934	0.990	0.895	0.937	0.987	0.900
mean		1.000	1.000	1.000	0.942	0.989	0.910	0.936	0.987	0.899
stdev		0.000	0.000	0.000	0.009	0.007	0.017	0.020	0.003	0.033

Table A3. Detailed classification results for female dorsal images

model	cv	training accuracy	training auc	training fl	validation accuracy	validation auc	validation fl	test accuracy	test auc	test fl
incv3	1	1.000	1.000	1.000	0.875	0.964	0.813	0.878	0.892	0.813
incv3	2	0.932	0.994	0.900	0.820	0.900	0.757	0.750	0.939	0.714
incv3	3	1.000	1.000	1.000	0.898	0.957	0.800	0.920	0.978	0.875
incv3	4	1.000	1.000	1.000	0.958	1.000	0.929	0.918	0.971	0.846
incv3	5	0.993	1.000	0.989	0.898	0.951	0.848	0.896	0.960	0.839
mean		0.985	0.999	0.978	0.890	0.954	0.829	0.872	0.948	0.817
stdev		0.030	0.003	0.044	0.050	0.036	0.064	0.071	0.034	0.062
resnet50	1	1.000	1.000	1.000	0.938	0.974	0.897	0.857	0.935	0.759
resnet50	2	1.000	1.000	1.000	0.940	0.977	0.914	0.979	0.994	0.968
resnet50	3	1.000	1.000	1.000	0.898	0.971	0.848	0.860	0.976	0.821
resnet50	4	1.000	1.000	1.000	0.938	0.976	0.909	0.898	0.984	0.839
resnet50	5	1.000	1.000	1.000	0.898	0.953	0.848	0.938	0.962	0.903
mean		1.000	1.000	1.000	0.922	0.970	0.883	0.906	0.970	0.858
stdev		0.000	0.000	0.000	0.022	0.010	0.033	0.052	0.023	0.080
incrnv2	1	1.000	1.000	1.000	0.958	0.996	0.938	0.918	0.978	0.867
incrnv2	2	1.000	1.000	1.000	0.880	0.948	0.833	0.979	1.000	0.968
incrnv2	3	1.000	1.000	1.000	0.959	0.990	0.929	0.940	0.974	0.909
incrnv2	4	1.000	1.000	1.000	0.938	0.992	0.897	0.959	0.982	0.929
incrnv2	5	0.925	1.000	0.893	0.776	0.949	0.732	0.729	0.933	0.683
mean		0.985	1.000	0.979	0.902	0.975	0.866	0.905	0.974	0.871
stdev		0.033	0.000	0.048	0.078	0.024	0.085	0.101	0.025	0.111

Table A4. Detailed classification results for female head lateral images

model	cv	training accuracy	training auc	training f1	validation accuracy	validation auc	validation f1	test accuracy	test auc	test f1
incv3	1	1.000	1.000	1.000	0.939	0.980	0.909	0.939	0.984	0.909
incv3	2	1.000	1.000	1.000	0.959	0.980	0.938	0.959	0.974	0.938
incv3	3	0.986	1.000	0.978	0.900	0.941	0.800	0.959	0.998	0.929
incv3	4	0.986	1.000	0.977	0.880	0.987	0.786	0.880	0.924	0.786
incv3	5	1.000	1.000	1.000	0.939	0.986	0.903	0.940	0.989	0.914
mean		0.995	1.000	0.991	0.923	0.975	0.867	0.935	0.974	0.895
stdev		0.007	0.000	0.012	0.032	0.019	0.069	0.033	0.029	0.062
resnet50	1	1.000	1.000	1.000	0.939	0.996	0.909	0.918	0.982	0.875
resnet50	2	0.987	1.000	0.978	0.898	0.978	0.815	0.898	0.982	0.800
resnet50	3	1.000	1.000	1.000	0.900	0.970	0.848	0.959	0.969	0.938
resnet50	4	1.000	1.000	1.000	0.900	0.957	0.857	0.920	0.955	0.882
resnet50	5	1.000	1.000	1.000	0.939	0.971	0.897	0.900	0.937	0.848
mean		0.997	1.000	0.996	0.915	0.974	0.865	0.919	0.965	0.869
stdev		0.006	0.000	0.010	0.022	0.014	0.038	0.025	0.019	0.050
incrv2	1	1.000	1.000	1.000	0.959	0.988	0.933	0.918	0.992	0.857
incrv2	2	1.000	1.000	1.000	0.959	0.996	0.933	0.939	0.982	0.897
incrv2	3	1.000	1.000	1.000	0.920	0.975	0.875	0.959	0.990	0.938
incrv2	4	1.000	1.000	1.000	0.940	0.987	0.909	0.880	0.958	0.813
incrv2	5	1.000	1.000	1.000	0.918	0.978	0.875	0.940	0.980	0.914
mean		1.000	1.000	1.000	0.939	0.985	0.905	0.927	0.981	0.884
stdev		0.000	0.000	0.000	0.020	0.008	0.029	0.030	0.014	0.049

A1.2 Nine-classes case

In all tables below the f1-score was weighted-averaged across classes.

Table A5. Detailed classification results for male dorsal images

model	cv	training accuracy	training f1	validation accuracy	validation f1	test accuracy	test f1
incv3	1	0.994	0.994	0.845	0.845	0.884	0.882
incv3	2	0.992	0.992	0.843	0.840	0.812	0.800
incv3	3	0.998	0.998	0.893	0.893	0.880	0.879
incv3	4	0.997	0.997	0.884	0.885	0.814	0.816
incv3	5	0.997	0.997	0.860	0.858	0.856	0.854
mean		0.996	0.996	0.865	0.864	0.849	0.846
stdev		0.003	0.003	0.023	0.024	0.035	0.037
resnet50	1	0.994	0.994	0.831	0.828	0.828	0.824
resnet50	2	0.995	0.995	0.880	0.878	0.789	0.784
resnet50	3	0.998	0.998	0.888	0.888	0.870	0.868
resnet50	4	0.988	0.987	0.856	0.857	0.837	0.838
resnet50	5	1.000	1.000	0.893	0.889	0.838	0.833
mean		0.995	0.995	0.870	0.868	0.832	0.829
stdev		0.005	0.005	0.026	0.026	0.029	0.030
incrnv2	1	1.000	1.000	0.878	0.876	0.893	0.890
incrnv2	2	1.000	1.000	0.870	0.872	0.831	0.830
incrnv2	3	0.995	0.995	0.856	0.853	0.829	0.826
incrnv2	4	1.000	1.000	0.894	0.891	0.842	0.839
incrnv2	5	0.997	0.997	0.879	0.876	0.870	0.867
mean		0.998	0.998	0.875	0.873	0.853	0.850
stdev		0.002	0.002	0.014	0.014	0.028	0.028

Table A6. Detailed classification results for male head lateral images

model	cv	training accuracy	training f1	validation accuracy	validation f1	test accuracy	test f1
incv3	1	0.998	0.998	0.823	0.817	0.798	0.794
incv3	2	0.997	0.997	0.858	0.859	0.814	0.817
incv3	3	1.000	1.000	0.886	0.885	0.849	0.848
incv3	4	1.000	1.000	0.868	0.870	0.854	0.858
incv3	5	0.997	0.997	0.794	0.789	0.850	0.850
mean		0.998	0.998	0.846	0.844	0.833	0.833
stdev		0.002	0.002	0.037	0.040	0.025	0.027
resnet50	1	0.995	0.995	0.836	0.835	0.780	0.780
resnet50	2	0.988	0.988	0.817	0.811	0.845	0.847
resnet50	3	1.000	1.000	0.868	0.865	0.822	0.819
resnet50	4	0.997	0.997	0.877	0.877	0.877	0.880
resnet50	5	1.000	1.000	0.861	0.861	0.877	0.873
mean		0.996	0.996	0.852	0.850	0.840	0.840
stdev		0.005	0.005	0.025	0.026	0.041	0.041
incrnv2	1	1.000	1.000	0.845	0.844	0.830	0.827
incrnv2	2	1.000	1.000	0.881	0.878	0.805	0.803
incrnv2	3	0.998	0.998	0.904	0.902	0.831	0.824
incrnv2	4	0.998	0.998	0.873	0.871	0.845	0.844
incrnv2	5	0.998	0.998	0.812	0.808	0.832	0.828
mean		0.999	0.999	0.863	0.861	0.828	0.825
stdev		0.001	0.001	0.036	0.036	0.015	0.014

Table A7. Detailed classification results for female dorsal images

model	cv	training accuracy	training f1	validation accuracy	validation f1	test accuracy	test f1
incv3	1	0.947	0.947	0.800	0.797	0.703	0.698
incv3	2	0.753	0.750	0.594	0.582	0.589	0.586
incv3	3	0.979	0.979	0.805	0.804	0.791	0.793
incv3	4	0.991	0.991	0.825	0.820	0.784	0.779
incv3	5	0.946	0.949	0.762	0.766	0.751	0.754
mean		0.923	0.923	0.757	0.754	0.724	0.722
stdev		0.097	0.099	0.094	0.098	0.083	0.084
resnet50	1	0.993	0.993	0.784	0.781	0.708	0.706
resnet50	2	0.991	0.991	0.829	0.830	0.784	0.786
resnet50	3	0.995	0.995	0.849	0.848	0.802	0.803
resnet50	4	0.995	0.995	0.746	0.743	0.751	0.750
resnet50	5	0.986	0.986	0.746	0.744	0.768	0.774
mean		0.992	0.992	0.791	0.789	0.763	0.764
stdev		0.004	0.004	0.047	0.048	0.036	0.038
incrnv2	1	1.000	1.000	0.800	0.795	0.778	0.776
incrnv2	2	0.980	0.980	0.733	0.735	0.692	0.688
incrnv2	3	0.746	0.740	0.595	0.567	0.567	0.536
incrnv2	4	1.000	1.000	0.831	0.828	0.805	0.801
incrnv2	5	0.998	0.998	0.822	0.817	0.847	0.837
mean		0.945	0.944	0.756	0.748	0.738	0.727
stdev		0.112	0.114	0.098	0.107	0.111	0.120

Table A8. Detailed classification results for female head lateral images

model	cv	training accuracy	training f1	validation accuracy	validation f1	test accuracy	test f1
incv3	1	0.991	0.991	0.812	0.805	0.770	0.743
incv3	2	0.983	0.983	0.768	0.748	0.770	0.750
incv3	3	0.998	0.998	0.861	0.859	0.779	0.774
incv3	4	1.000	1.000	0.796	0.789	0.799	0.795
incv3	5	1.000	1.000	0.843	0.837	0.796	0.793
mean		0.994	0.994	0.816	0.808	0.783	0.771
stdev		0.008	0.008	0.037	0.043	0.014	0.024
resnet50	1	0.988	0.988	0.796	0.799	0.754	0.756
resnet50	2	0.995	0.995	0.768	0.749	0.801	0.790
resnet50	3	0.991	0.991	0.840	0.842	0.721	0.717
resnet50	4	1.000	1.000	0.801	0.791	0.804	0.803
resnet50	5	0.984	0.984	0.832	0.826	0.733	0.729
mean		0.992	0.992	0.808	0.801	0.763	0.759
stdev		0.006	0.006	0.029	0.036	0.038	0.037
incrnv2	1	0.998	0.998	0.843	0.844	0.832	0.832
incrnv2	2	0.998	0.998	0.800	0.791	0.853	0.841
incrnv2	3	0.995	0.995	0.871	0.881	0.742	0.757
incrnv2	4	0.993	0.993	0.812	0.806	0.825	0.824
incrnv2	5	0.997	0.997	0.827	0.825	0.770	0.758
mean		0.996	0.996	0.831	0.829	0.804	0.802
stdev		0.002	0.002	0.028	0.035	0.047	0.041