






Image-based taxonomic classification of bulk insect biodiversity samples using deep learning and domain adaptation

Tomochika Fujisawa¹  | Víctor Noguerales^{2,3}  | Emmanouil Meramveliotakis²  |
Anna Papadopoulou²  | Alfried P. Vogler^{4,5} 

¹The Center for Data Science Education and Research, Shiga University, Hikone, Japan

²Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus

³Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), Tenerife, Spain

⁴Department of Life Sciences, Natural History Museum, London, UK

⁵Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, UK

Correspondence

Tomochika Fujisawa, The Center for Data Science Education and Research, Shiga University, 1-1-1 Banba, Hikone, Shiga 522-8522, Japan.
Email: t.fujisawa05@gmail.com

Víctor Noguerales, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), Astrofísico Francisco Sánchez 3, 38206, San Cristobal de La Laguna, Tenerife, Spain.
Email: victor.noguerales@csic.es

Funding information

Agencia Estatal de Investigación; European Union's Horizon 2020, Grant/Award Number: 810729; Juan de la Cierva-Formación, Grant/Award Number: FJC2018-035611-I; JSPS KAKENHI, Grant/Award Number: 20K06824

Abstract

Complex bulk samples of insects from biodiversity surveys present a challenge for taxonomic identification, which could be overcome by high-throughput imaging combined with machine learning for rapid classification of specimens. These procedures require that taxonomic labels from an existing source data set are used for model training and prediction of an unknown target sample. However, such transfer learning may be problematic for the study of new samples not previously encountered in an image set, for example, from unexplored ecosystems, and require methods of domain adaptation that reduce the differences in the feature distribution of the source and target domains (training and test sets). We assessed the efficiency of domain adaptation for family-level classification of bulk samples of Coleoptera, as a critical first step in the characterization of biodiversity samples. Neural network models trained with images from a global database of Coleoptera were applied to a biodiversity sample from understudied forests in Cyprus as the target. Within-dataset classification accuracy reached 98% and depended on the number and quality of training images, and on dataset complexity. The accuracy of between-datasets predictions (across disparate source–target pairs that do not share any species or genera) was at most 82% and depended greatly on the standardization of the imaging procedure. An algorithm for domain adaptation, domain adversarial training of neural networks (DANN), significantly improved the prediction performance of models trained by non-standardized, low-quality images. Our findings demonstrate that existing databases can be used to train models and successfully classify images from unexplored biota, but the imaging conditions and classification algorithms need careful consideration.

KEYWORDS

biodiversity assessment, bulk sample, coleoptera, convolutional neural network, domain adaptation, image classification, machine learning

INTRODUCTION

Biological identifications increasingly rely on machine learning algorithms that use photographic images to place unidentified specimens into a taxonomic classification. As these methods are proving to

Tomochika Fujisawa and Víctor Noguerales contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Systematic Entomology* published by John Wiley & Sons Ltd on behalf of Royal Entomological Society.

be very powerful especially for identification of the species-rich and morphologically diverse insects, it is now possible to place a specimen with high confidence against curated image libraries, for example, those obtained from pinned museum collections (Buschbacher et al., 2020; Hansen et al., 2020). With the rapid increase of such images, machine learning can greatly increase the capacity for species identification without putting demand on scarce taxonomy experts (Høye et al., 2021; Valan et al., 2019). The methodology, therefore, is likely to play a major role in the taxonomic endeavour in future, and deep learning potentially can have similar impacts on the practice of taxonomy as the revolution of DNA barcoding and metabarcoding some 20 years ago, or it could work in concert with these molecular approaches (Høye et al., 2021; Wühl et al., 2022; Yang et al., 2022). However, the true potential and possible limitations of algorithmic methods for exploiting the information contained in specimen images remain to be established, as the various applications and choice of machine learning algorithms continue to be refined (Romero et al., 2020; Valan et al., 2019).

The greatest challenge for modern taxonomy probably is the study of highly diverse and poorly studied biotas and geographic regions around the world, harbouring many undescribed species (Costello et al., 2013). In particular, in studies of insect diversity, such as those from tropical forest canopy or the soil, huge numbers of specimens are collected and subsequently need to be classified and counted as part of ecological and environmental studies (Novotny et al., 2007; Caruso et al., 2019). In these circumstances, specimens are often assigned to high taxonomic ranks at order and family levels (Karlsson et al., 2020), for example, for broad ecological comparisons (Stork & Grimbacher, 2006) and ecological status assessment using bulk-sample specimens in freshwater ecosystems (Escribano et al., 2018). Thus, despite the lack of taxonomic resolution, family-level assignments are a critical first-step to characterize biodiversity samples and demand rarely available broad knowledge of insects across taxonomic groups and geographic regions. Imaging of these specimens is comparatively fast with the help of recently described automated imagers (Ärje et al., 2020; Wühl et al., 2022) or by taking high resolution images of large sets of specimens in a single photo, which can then be cropped to represent single individuals for subsequent classification (Hudson et al., 2015; Bian et al., 2022). Automated classification based on these images would remove the need for manual identification by taxonomic experts who individually can handle only a small portion of the diversity spectrum usually encountered in such studies (Basset et al., 2012), and thus may help to provide rapid assessment of threatened insect assemblages, where speed is a priority.

In machine learning, images are classified against a set of defined objects, for example, the images of a particular taxonomic group. A model trained to separate the types of images in this source is used to classify unlabelled objects in the target, such as an unknown set of specimens. Most recent studies used convolutional neural networks (CNN, LeCun et al., 2015) for the task of image classification. Because of the lack of images for training the full parameters of a CNN model, approaches like fine-tuning of the existing CNN (Ärje et al., 2020) or

feature transfer from the pre-trained CNN (Valan et al., 2019) are commonly used in biodiversity studies, following the successful applications of pre-trained CNN outputs as generic image features (Donahue et al., 2013; Razavian et al., 2014). These methods of transfer learning (sensu Valan et al., 2019; see Table S1 for a detailed terminology) have already shown great power in taxon annotation of insect specimens, and in some cases, surpass the capabilities of trained taxonomists (Valan et al., 2021).

Yet, applications of image classification algorithms for insect biodiversity research have mainly been limited to narrow tasks and specific target sets, such as pinned museum specimens (Hansen et al., 2020; Valan et al., 2019), aligned body parts (Buschbacher et al., 2020; Klasen et al., 2022), or small target groups of a few species (Ärje et al., 2020; Popkov et al., 2022). In most of these studies, the unlabeled (target) set is from the same dataset, that is, the target taxa at species or higher hierarchical levels are included in the training set. However, as hitherto unsampled specimen sets are included, the feature space of source and target domains no longer has similar distributions. Thus, aligning the disparity between domains requires a trained model that can be generalized across the entire feature space of the domains, using procedures of 'domain adaptation' (e.g., Pan & Yang, 2010; Farahani et al., 2020; see Table S1). Although methods for domain adaptation have been successfully applied to fields such as medical image classification (Guan & Liu, 2021), they may also be useful for analysis of biodiversity samples and the classification of insect specimens from unexplored areas whose components are unlikely to be present in the training set.

Building an image-based classification system may be further complicated by several factors affecting the feature distribution of source and target datasets. Capture bias is a well-established problem in machine learning, as objects appear in different contexts (location, lighting, background, etc.) or are taken on different imaging devices. Images of insects may be from collection specimens taken in fairly standardized positions and lighting conditions (Hansen et al., 2020; Valan et al., 2019), or may be obtained directly from bulk samples and photographed either singly (Raitoharju et al., 2018; Valan et al., 2019; Wühl et al., 2022) or cropped from large-field composite images (Buschbacher et al., 2020; Hansen et al., 2020). Images thus display different aspects of the specimens and differ in illumination and magnification, which affects the recognition of key features (Ärje et al., 2020; Raitoharju et al., 2018). The performance of a model trained in one dataset can be compromised if this deviation of a prediction target from the training source is not controlled correctly (Torralba & Efros, 2011; Tommasi et al., 2017), and such performance reduction has already been reported in applications for biodiversity research (Knyshov et al., 2021; Popkov et al., 2022).

Other issues are unrelated to differences in image acquisition, but result from the biases of defining the semantic categories or classes recognized in the source and target domains (Tommasi et al., 2017). Such 'category bias' may arise from inconsistent labelling, either due to the application of different taxon concepts used for classifying species and higher taxa, or due to specimen misidentification. The resulting noisy or incorrect data labels then reduce the effectiveness of the

model. In addition, in particular, in higher taxonomic categories, the same name is assigned to visually different images due to the distributional shift of subclasses (e.g., different genera representing a family in the source and target). Furthermore, in general cross-dataset applications, the model can encounter a category which is missing in the source training data, for example, a new family may be present. The treatment of such anomalous (or 'out-of-distribution'; Tabak et al., 2019; see Table S1) samples affects the reliability of the biodiversity assessment. As more variation is encountered, to fully learn the structure of the data, the model should scale with the size and complexity of the training data.

In practice, due to these problems of intra-class variability and the inconsistencies of the photographs, the success of deep learning in taxonomy to date has been in situations where a bespoke image library is available that holds a narrow representation of the query taxa and images under the same aspect and imaging conditions (brightness, angle, magnification, etc.; Buschbacher et al., 2020; Valan et al., 2021). In addition, the performance evaluation in these studies often has been limited to the training-testing procedure within a single dataset, and the generalization capability of the models across datasets was not explicitly examined. The utility of these methods remains largely untested in the application to samples from poorly characterized species, as those from previously unseen bulk insect samples in unexplored areas. Ideally, such samples would be identifiable against images drawn from other sources, for example, an image database of well characterized regional communities and taxa obtained elsewhere, although minimizing the adverse effects of biases in these datasets.

Here, we use deep learning approaches for classification of insects based on bulk-sample images from high-throughput biodiversity surveys, testing the possibility of domain transfer between unrelated image sets. We characterize bulk samples of poorly known communities of Coleoptera (beetles) collected in different sites around the world, whose high species diversity and complex morphological variation provide a challenging, but realistic situation for machine-based classification. The aim was a classification at the taxonomic level of family, which is a frequent goal of initial identification in world-wide biodiversity surveys where results may be used for subsequent specimen counts, assignment to functional groups, or further in-depth taxonomic identification by specialists. Using a specimen database from sites of various geographic origin globally identified at family level, we attempt the classification of specimens in a geographically and ecologically disparate community (Cyprus). We use this setup to address the question about the transferability of identifications across communities from different habitats and continents, that is, when the input subclass (species and genera within a family) is not present in the training data. Various parameters are tested that may affect the prediction accuracy, including: (i) the size of the training set; (ii) the complexity of the training set, which may be affected by the level of intra-class variability, noise from misidentifications, or the presence of out-of-distribution samples; and (iii) the quality of images, for example, the resolution of the image using standard macrophotography versus high-resolution stacking technology. The error from these factors may be reduced by the use of advanced methods for

domain adaptation. We here apply one such method, the domain adversarial neural network (DANN) algorithm, which includes unlabeled images from the target dataset in its training process to improve the target prediction. As we show, the use of deep learning with the specific domain adaptation algorithm is a powerful approach for classifying unknown samples but the prediction success depends on the composition of the training set and may vary between classes (i.e., some beetle families are more easily predicted).

MATERIALS AND METHODS

Sample collection and taxon selection

As the target for classification, we used a collection of leaf-litter bulk samples from a total of 46 sites distributed across five forest habitats of the Troodos mountain range of Cyprus (Figure 1). These samples were processed as described by Nogueras et al. (2021) to extract bulk Coleoptera specimens from the substrate using a Berlese apparatus. During bulk-sample processing, a subset of individual specimens, representing all different morphospecies encountered in the samples, were separated and processed alongside the remainder of the bulk samples. The two sets of samples (bulk samples vs. single specimens) were preserved in 100% ethanol and subsequently photographed following two different imaging protocols (L_H and L_L , respectively, see below). For more details on soil sampling and habitat descriptions, see Arribas et al. (2016) and Nogueras et al. (2021), respectively. During sample processing and imaging, the most common families/subfamilies, with 5 or more photographs per taxonomic rank and dataset, were identified and used for downstream analysis. The chosen families were: Brentidae, Carabidae, Chrysomelidae, Cryptophagidae, Curculionidae, Latridiidae, Leiodidae, Melyridae, Ptiliidae, Staphylinidae: Scaphidiinae, Staphylinidae (excluding Scaphidiinae) and Tenebrionidae (Table S2).

Image data acquisition

Local high quality (L_H) dataset

Bulk samples were air-dried and specimens placed at regular distances onto filter paper in a Petri dish. In cases of large disparity in body size, we split the bulk samples into different size categories which were separately photographed in order to improve the focus and resolution across all specimens regardless of their body size. As much as possible, specimens were positioned for photography in dorsal view.

Bulk-sample photographs were taken using a Zeiss AXIO Zoom.V16 Stereo Zoom Microscope equipped with a Zeiss AxioCam HRc (High Resolution 13 Megapixels Colour Microscope) camera at the Imaging and Analysis Centre at the Natural History Museum (NHM) in London, United Kingdom. This instrument has a motorized focus drive and motorized stage for generating large high-resolution images by dividing the field into regular tile-images that are subsequently xyz stitched. Depending on the sample size, photographs were taken by dividing them

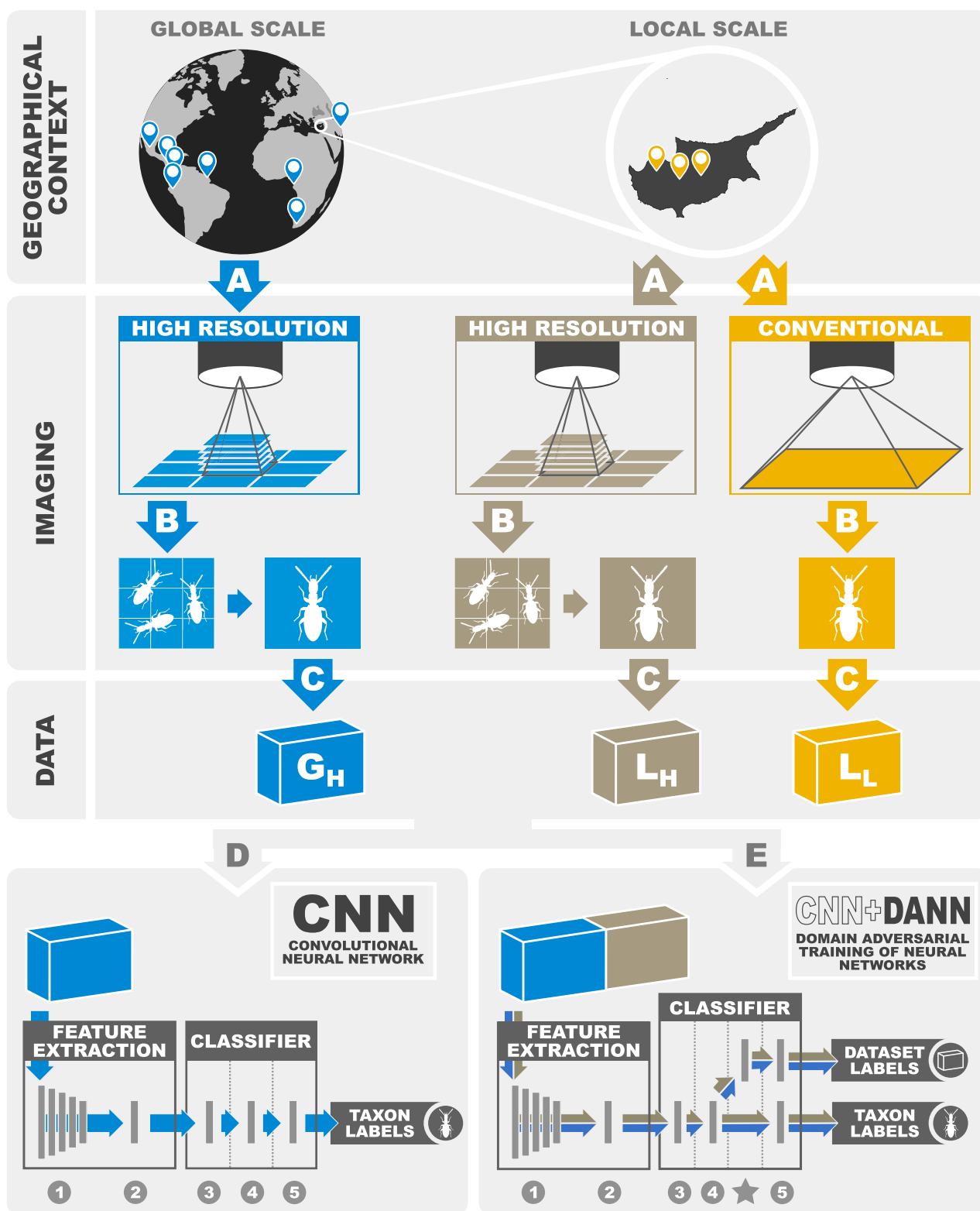


FIGURE 1 Schematic diagrams summarizing the experimental workflow of the study, depicting the geographical context and the different imaging procedures for generating the three image datasets: G_H , global high quality; L_H , local high quality; L_L , local low quality. Taxon classification was performed using two alternative deep learning algorithms: Convolutional neural network (CNN) and domain adversarial neural network (DANN). For more details on algorithm-specific architectures, see Figure S1.

into 16–64 tiles, each with 25–30 slices (z-stacks) using the Zeiss NEO 2 Blue Edition software. We rendered z-stack images with the Helicon Focus v.5.3.14 software (<https://www.heliconsoft.com>) using the

pyramid-based algorithm ('Method C') and default parameters. Focus stacking was also performed using the depth-map algorithm ('Method B') in Helicon Focus with a radius value of 8 and a smoothing parameter

of 4, yielding qualitatively similar images to the former method. Only photos from 'Method C' were used for downstream analyses.

Finally, we manually cropped individual specimen photos from the bulk-sample images using INSELECT v.0.1.35 software (Hudson et al., 2015). After some minor corrections of bounding edges, cropped single-specimen images were exported and taxonomically identified at the family/subfamily level by the authors. Only whole-bodied specimens were considered for further analyses. The cropped images were resized to 255×255 pixels for subsequent classification tasks. When an image was not an exact square, the edges were padded using the average pixel value of the outermost portions of the image to enforce a square shape.

The individual frames cropped from the bulk samples were denoted the *Local High Quality* (L_H) data set, referring to the fact that they were obtained from a local area and thus represent a small taxonomically confined set, and taken at high image resolution. The L_H dataset represented the best-case scenario, where high-resolution training images of local samples are obtained under controlled conditions with high-performance imaging equipment. This set was the primary target in measuring the success of transfer learning.

Local low quality (L_L) dataset

A subset of single specimens (taken from the bulk samples) were individually photographed using a conventional stereoscope NIKON SMZ1270i equipped with a NIKON DS-Fi3 Microscope Camera (5.9 megapixels) controlled by the NIKON DS-L4 v.1.5.0.3 control unit. These images were denoted *Local Low Quality* (L_L) dataset. These photographs were intended to represent a more realistic scenario of local specimens being photographed during field sampling and sample sorting in local laboratory facilities using conventional instruments, and were used to address the question about 'capture bias', that is, the effect of imaging conditions on classification accuracy.

Global high quality (G_H) dataset

We also obtained a wider sample of images from a global catalogue of Coleoptera specimens available at <https://www.flickr.com/photos/site-100/>. These images had been obtained from local sampling campaigns at 11 sites throughout Central America, Africa and Southeastern Asia (see Table S3) and photographed in bulk using the Zeiss AXIO Zoom, as described above, although others were individually taken at high-resolution on a single lens reflex (SLR) camera (Canon EOS 500D) and macro lens (Canon MP-E 65 mm f/2.8 1-5x Macro). Helicon Focus software was used to render z-stack images, as described above. This dataset was denoted the *Global High Quality* (G_H) dataset. For each of the selected families, all specimen photographs available for the respective sites were used. Relative numbers of available specimens per family were usually correlated across sites, with greatest numbers in Staphylinidae. The numbers of images in the three data sets are shown in Table S2. The G_H dataset mainly consists of samples from tropical

forest interception traps and leaf litter, and does not share lower taxonomic groups with the target dataset sampled from Mediterranean forest soils. These collections were the source for the test of domain adaptation protocols applied to the unknown Cypriot target.

Image classification with neural network (NN)

Feature transfer and neural network classifier

We employed the strategy of feature transfer from the pre-trained convolutional neural network (CNN) proposed by Valan et al. (2019). We chose the outputs of the fifth convolutional block of the VGG19 model after 2-dimensional average pooling as a set of features for an image, based on the results of Valan et al. (2019) and our pilot analyses. These 512-dimensional image features were used for the classification with a neural network classifier.

The neural network classifier consisted of two fully connected (FC) layers with ReLU activation and a softmax output layer (Figures 1 and S1). The dropout was applied after the FC layers with a dropout rate of 0.6. The neural network was trained with the stochastic gradient descent algorithm with the softmax cross-entropy loss for 300 epochs. We used a batch size of 10 and a fixed learning rate of 0.01, and the convergence of loss was visually assessed. The numbers of units in the two FC layers (512 and 256 for the first and second FC layers, respectively) and the dropout rate were determined by five-fold cross-validation with a random 200 images of the G_H dataset, and these hyperparameters were used throughout all classification tasks in this study.

Metrics for prediction accuracy

We evaluated the performance of the models with the following metrics throughout the subsequent classification experiments. The accuracy of the prediction was measured as the proportion of successful predictions in the test set, $Acc = \frac{1}{n} \sum_{i=1}^n [\hat{y}_i = y_i]$, where \hat{y}_i is the predicted class of the i -th image, y_i , the true class and $[\hat{y}_i = y_i]$ is 1 if $\hat{y}_i = y_i$ and 0 otherwise.

The classification performance for each class was measured by the multiclass recall rate, multiclass precision and the F1-score (see Table S1). Recall rate of class c is defined as a proportion of correct predictions of c out of the actual number of images of c ,

$$Recall_c = \frac{\sum_{i=1}^n [\hat{y}_i = y_i] [y_i = c]}{\sum_{i=1}^n [y_i = c]}.$$

Multiclass precision is defined as a proportion of correct predictions of c out of the number of images predicted as c ,

$$Precision_c = \frac{\sum_{i=1}^n [\hat{y}_i = y_i] [\hat{y}_i = c]}{\sum_{i=1}^n [\hat{y}_i = c]}.$$

The F1-score is the harmonic mean of the multiclass recall rate and precision. Thus, the recall rate is interpreted as the fraction of images of a class present in the sample that are correctly selected, while precision quantifies the fraction of the images predicted as members of a class that are actually correct. The F1-score represents the overall performance of a classifier with respect to these two measures.

We evaluated the transferability of learning by measuring the reduction of accuracy when a model trained with a source data set predicts target images. The target accuracy, Acc_T , was measured as the proportion of successful predictions of the target images (see Table S1). The baseline accuracy within the source dataset, Acc_S , measured by the within-dataset classification was compared with Acc_T . The accuracy reduction, $\Delta Acc(S, T) = Acc_S - Acc_T$, was recorded as a measure of transferability between the datasets. High ΔAcc indicates large reduction of accuracy, hence difficulty in transfer.

Divergence between the source and target datasets was measured with a dataset classification error. A linear support vector machine (SVM) was trained to classify images to the source or target dataset with the features of 200 randomly selected images from both datasets. Conversely to the above analyses, here the model was trained to classify datasets instead of taxa. Then, a classification error of the SVM, $\varepsilon_{source-target}$, was measured as a proportion of incorrect predictions of 200 test images sampled from the two datasets. An intuitive interpretation of this measure is that the dataset classification task is harder when the feature distributions between two datasets are more similar. Therefore, a large classification error indicates high similarity between source and target datasets. This approach is commonly used to measure the dataset bias (Tommasi et al., 2017).

Within-dataset classification

To evaluate the baseline performance, Acc_S , of the CNN model, we first conducted bulk image classification within datasets (assessing the effect of intra-class variability). This was performed by testing the number of training images on prediction accuracy, whereby the CNN model was trained with N images randomly selected from the dataset and predicted the class (family label) of n test images randomly selected from the rest. N ranged between 100 and 700 for L_H (with intervals of 100 images), between 50 and 250 for L_L (with intervals of 50 images) and between 100 and 900 for G_H (with intervals of 100 images). The number of test images n was set to 200 for L_H and G_H , and 50 for L_L due to the small size of the dataset. To evaluate the consistency of prediction accuracy, 10 replicates were generated for each scenario of N images. The effects of the number of images and difference of prediction accuracy between datasets were assessed by a linear regression model.

Between-datasets classification

For the between-dataset prediction, the CNN model was trained with a source dataset to predict images from a different target dataset. The NN was trained with N images randomly selected from the source dataset, which was then used to predict all images of the target dataset and Acc_T and ΔAcc were measured. We ran the above procedures for three source–target pairs (training dataset \rightarrow predicted dataset), $G_H \rightarrow L_H$, $G_H \rightarrow L_L$ and $L_L \rightarrow L_H$. These settings simulate two alternative scenarios: (i) a global image database is used to predict local samples ($G_H \rightarrow L_H$ and $G_H \rightarrow L_L$) and (ii) conventional images, as those

representing single-specimen photographs by local taxonomists, are used to predict local high-resolution images ($L_L \rightarrow L_H$).

Between-datasets classification with domain adversarial training

In addition to the standard CNN setups described above, we employed the domain adversarial training of neural networks (DANN, Ganin et al., 2016) which incorporates a certain portion of the unknown targets in the model. The DANN model jointly predicts the class (family label) of the source images and the dataset (domain) of all input images (as in the previous section) by adding layers for the dataset classification to the classifier (Figure S1). The training procedure then optimizes the model parameters in the shared part of the network to not only minimize the loss of the label classifier (taxon prediction) but at the same time to maximize the loss of the domain classifier (dataset prediction). This adversarial training procedure optimizes shared intermediate features to be invariant between the two domains, and hence the model can generalize across them, which potentially improves the accuracy in target predictions. In this study, a softmax layer with binary cross entropy loss was added as a dataset classifier to the NN after the second FC layer. The regularization parameter, λ , which controls the relative importance of the two classifiers, was set to $\lambda = 0.1, 0.5$ and 1.0 , and the best performing results ($\lambda = 0.1$) were reported.

The performance of the DANN method was measured with procedures similar to those in the previous section. A mixed set of images of size N was randomly selected from target and source datasets, and training was done using taxon labels from the source images and dataset labels for all images. Next, 400 mixed test images were predicted, and their Acc_S , Acc_T and ΔAcc were recorded. We applied the DANN to the three pairs from the previous section. The total number of images N ranged between 300 and 800 for $L_L \rightarrow L_H$, 400 and 1400 for $G_H \rightarrow L_H$, and 300 and 1000 for $G_H \rightarrow L_L$. The proportions of source images were 0.3, 0.67 and 0.83 for $L_L \rightarrow L_H$, $G_H \rightarrow L_H$ and $G_H \rightarrow L_L$, respectively, which yielded training images from the source similar in number to the other training setups. The effect of DANN on target accuracy was tested using linear regression with the model type and the number of images as explanatory variables. Models of neural networks were implemented in Python with Keras 2.5.0 (<https://keras.io>) and TensorFlow 2.5.0 (<https://www.tensorflow.org>) libraries, and all statistical analyses were conducted with R 4.1.0 (R Core Team, 2021).

RESULTS

Performance of within-dataset classification

Effects of datasets and the number of images

The accuracy of within-dataset classification and the effect of the number of training images varied among datasets. The accuracy for the L_H samples of specimens collected from Cyprus generally

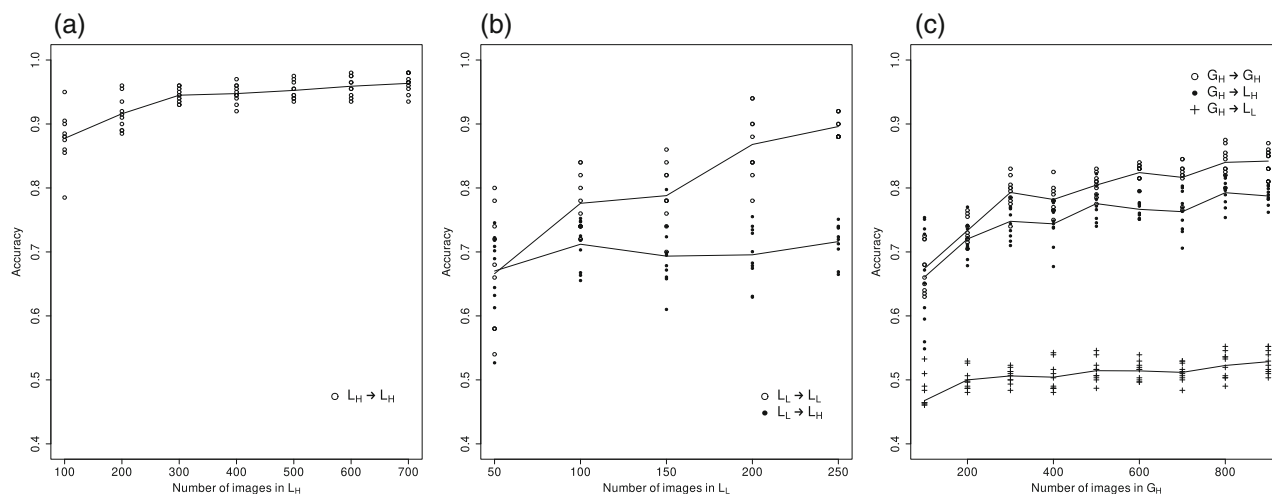


FIGURE 2 Effect of increasing the number of training images on prediction accuracy. Training the convolutional neural network (CNN) on a subset of images and prediction of the class (family label) of images. (a) *Local high quality* (L_H) images for training and predicting the class of L_H images, (b) *local low quality* (L_L) images for training and predicting the class of either L_L or L_H images and (c) *global high quality* (G_H) images for training and predicting the class of either L_L , L_H or G_H images.

improved with an increasing number of training images and reached an average of 96% with 700 images (Figure 2a). The maximum classification accuracy for the L_H was 98%.

The within-dataset classification accuracy of the L_L images, taken by a conventional stereoscope and camera, was generally lower compared to the L_H dataset (9.1% lower for L_L , linear regression p -value, $p < 0.001$). The accuracy increased monotonically with the increasing number of images and reached an average of 89% with 250 images (Figure 2b). As expected, the within-dataset classification accuracy of the G_H images obtained from diverse sites around the world was significantly lower (16% lower for G_H , $p < 0.001$) compared to the L_H obtained from the single area of Troodos. The improvement of accuracy was slower than for the other datasets, and the average accuracy was 84% with the maximum number of 900 images (Figure 2c), consistent with the greater heterogeneity of the global set. Loss and accuracy development during training of models are reported in Figures S2, S3.

Performance of between-dataset classification

The accuracy of cross-dataset predictions was first assessed in regard to the effect of image quality. When the L_L images were used to train the NN and then to predict the L_H images, the accuracy remained largely constant at 71% for 250 images (Figure 2b). The accuracy reduction ΔAcc , that is, the reduction in success of predictions compared to the predictions expected from within-dataset classification, rapidly increased with the number of images (Spearman $\rho = 0.72$, $p < 0.001$), indicating that the training with L_L images did not improve the prediction of the L_H images (Figure 3).

Next, we considered the critical question about the power of the global dataset to predict the local data, using the G_H and the L_H as a source–target pair. The prediction accuracy for this comparison was

close to the within- G_H predictions, with the average accuracy being 79% and the maximum 82% with 900 images (Figure 2c), indicating that the local set from the Cyprus collection (L_H) behaved in a similar way as the other local sets contributing to the G_H dataset. The accuracy reduction from G_H to L_H was on average 0.04 and remained almost constant after 300 images ($\rho = 0.13$, $p = 0.28$, Figure 3). The power of the G_H dataset required the high image quality exhibited by the target (L_H); when the G_H -trained model was used to predict the L_L images, the accuracy was significantly lower (Figure 2c). This was also evident from the increased accuracy reduction with increased number of images; whereas, the $G_H \rightarrow G_H$ predictions improved with more images, the $G_H \rightarrow L_L$ predictions did not (Figure 3). The dataset classification errors ($\epsilon_{source-target}$) were 0.20 ($G_H \rightarrow L_H$), 0.06 ($G_H \rightarrow L_L$) and 0.01 ($L_H \rightarrow L_L$), indicating high similarity between the G_H and L_H images and the distinctiveness of the L_L .

The performance of the domain adversarial training

The DANN significantly improved the target accuracy of the $L_L \rightarrow L_H$ prediction, which involves images from the different photographic setups (Figure 4a,b). A linear regression model showed that the target accuracy increased by 6.2% ($p < 0.001$, Figure 4b) and the accuracy reduction decreased by 0.060 when the DANN model was used with labelled L_L and unlabelled L_H images (Figure 3). The average target accuracy was 79% with 200 labelled L_L images and 400 unlabelled L_H images (Figure 4b), approaching the same level of accuracy as $G_H \rightarrow L_H$ predictions.

On the contrary, the DANN did not improve the target accuracy when the G_H was used as a source dataset (Figure S4 and S5). The $G_H \rightarrow L_H$ target accuracy was on average 0.75 with 940 labelled G_H images and 460 unlabelled L_H images (in total 1400 images), and overall target accuracy was significantly lower than the between-dataset

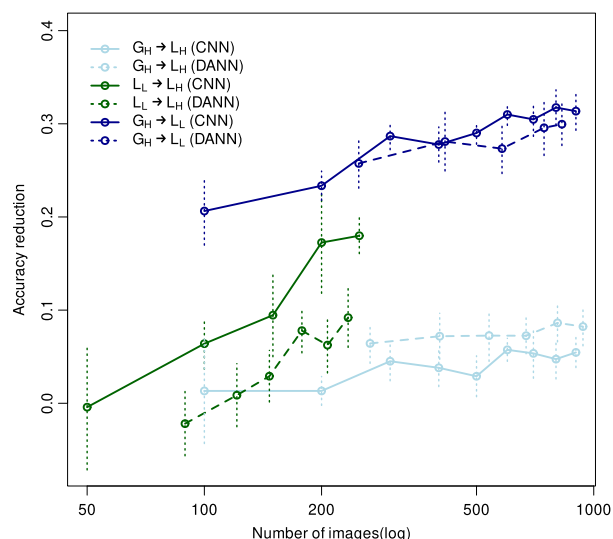


FIGURE 3 The effect of increasing numbers of training images on the accuracy reduction in across-dataset predictions. Subsets of randomly selected images of one dataset are used for training and predicting the class (family label) of another set, as indicated by different colours. Lines in light blue refer to the comparison involving tests of locality, that is, when using global high quality (G_H) images for training and predicting the class of local high quality (L_H) images. Lines in green refer to comparisons involving tests of image quality, that is, when using local low quality (L_L) images for training and predicting the class of local high quality (L_H) images. Lines in dark blue refer to comparisons involving differences in both locality and image quality, that is, when using global high quality (G_H) images for training and predicting the class of local high quality (L_L) images. The x-axis representing the number of training images is on a logarithmic scale. The vertical dotted bars indicate 95% confidence interval of the average accuracy reduction. Higher accuracy reduction indicates a worse performance on prediction compared to the within-dataset prediction accuracy. The solid and dashed lines represent results of the convolutional neural network (CNN) and domain adversarial neural network (DANN), respectively. Note that only the L_L to L_H prediction accuracy improved with the use of DANN.

predictions by the plain NN model (2.2% reduction for DANN, $p = 0.0002$, Figure S4). In the $G_H \rightarrow L_L$ prediction, a similar trend was observed (Figure S5) and the target accuracy was not significantly different from the NN (0.015% reduction for DANN, $p = 0.98$). Loss and accuracy development during training of models are reported in Figures S2, S3.

Classification error

Classification error was visualized as a scaled confusion matrix. Starting with a trial for a within-dataset analysis with 400 training images in the L_H random sampling showed that the large taxonomic groups were correctly classified in most cases (Table S4). For example, four families (Carabidae, Curculionidae, Ptiliidae and Staphylinidae) were classified with more than 95% recall rate, although the remaining taxa had widely different recall rates ranging from 0% to 82% (Figure 5a). In the

extreme case of the family Melyridae, with the lowest number of available images ($n = 5$), no images were predicted correctly (Figure 5a). When a taxon had >50 images, its recall rate and precision approached 1.0 (Figure 5a,c). The F1-scores showed a similar pattern, that is, for those images that were called to be members of a taxon, these predictions were generally correct (Figure 5e). Class-wise recall rates and F1-scores showed a strong positive correlation with the number of images (recall rates: $\rho = 0.81$, $p = 0.0014$; F1-scores: $\rho = 0.85$, $p = 0.0005$; Figure 5a,e). The effect of the number of images on the class-wise precision was also positive, but slightly weaker ($\rho = 0.41$, $p = 0.187$, Figure 5c). Failed predictions included ventral views of insect bodies, specimens with missing body parts or multiple specimens in a single image (see Figure S6). Apart from these irregular images, most failed predictions were for taxa represented by <20 images (Figure 5a,c,e). Prediction probabilities for the successful predictions (average 0.98) were overall higher than for the failed predictions (average 0.79, Figure 6a), when using the L_H dataset with 400 training images.

For the between-dataset analysis, a confusion matrix of the $G_H \rightarrow L_H$ prediction trained on 800 images equally showed a high accuracy of predictions (Table S5). Misclassification mostly affected morphologically similar taxa, for example, the reciprocal confusion of Brentidae and Curculionidae (Table S5). Chrysomelidae, Curculionidae and Staphylinidae had recall rates >0.90 (Figure 5b), but more taxa were incorrectly classified than in the case of the $L_H \rightarrow L_H$ prediction. No image of Leiodidae and Scaphidiinae, with the available training images <50 , was predicted correctly (Figure 5b).

The success of the class-wise recall rates was strongly correlated with the number of images in the source dataset ($\rho = 0.77$, $p = 0.0036$, Figure 5b). Three taxa with >300 images had recall rates >0.95 , although the taxa with <40 images had recall rates <0.4 (Figure 5b). The effect of the number of images on the class-wise precision and F1-score was also positive, but the effects were not significant (class-wise precision: $\rho = 0.16$, $p = 0.618$; F1-score: $\rho = 0.42$, $p = 0.171$; Figure 5d,f). Surprisingly, the F1-scores were greatly reduced relative to the recall score for the Chrysomelidae, indicating the precision of the prediction was low even if the recall was high (Figure 5d,f), that is, the true Chrysomelidae were correctly classified, but many other taxa were incorrectly classified as Chrysomelidae.

Prediction probabilities and out-of-distribution samples

In order to test the effect of the presence of unknown inputs (out-of-distribution samples) on the classification, we first used an L_H -trained model to predict the class of 16 L_H images belonging to eight families/subfamilies, Coccinellidae, Elateridae, Endomychidae, Hydrophilidae, Laemophloeidae, Phalacridae, Scarabaeidae and Scydmaeninae, which were not present in the training data, but were present in the target sample (Cyprus) in small numbers. For these images, the (incorrect) prediction probabilities were also lower on average than for the

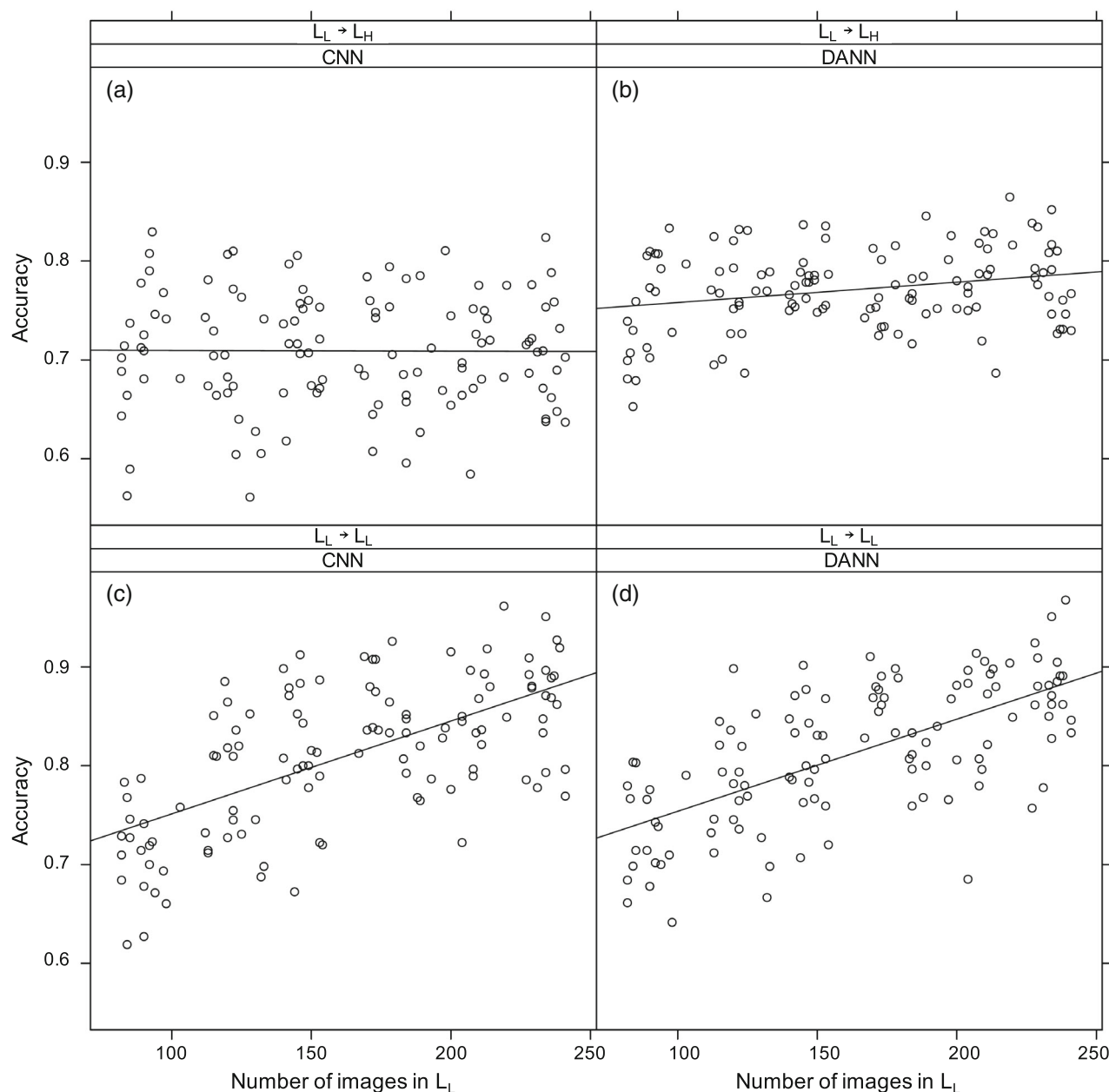


FIGURE 4 Effect of the number of images on prediction accuracy of the convolutional neural network (CNN, panels a and c) and the domain adversarial neural network (DANN, panels b and d) training for the *local low quality* (L_L) and *local high quality* (L_H) images. Top panels (a and b) represent between-dataset predictions ($L_L \rightarrow L_H$) and bottom panels (c and d) indicate within-dataset predictions ($L_L \rightarrow L_L$). Solid lines represent regression lines between the number of images and accuracy. For both between- and within-dataset predictions, models using DANN were trained with a mixed set of randomly selected images from the L_L and L_H datasets. For other dataset comparisons, see Figures S3 and S4.

successful predictions (average 0.83, Figure 6a). However, four samples were predicted with high probabilities of >0.95 , including three images of Coccinellidae, Hydrophilidae and Phalacridae that were classified as Ptiliidae (Figure 6a).

A similar test was also conducted for the between-dataset classifications by using a G_H -trained model to predict L_H images. Similar to the within-dataset classification, average prediction probabilities of successful predictions (0.98) were consistently higher than the failed predictions (0.84) and out-of-distribution samples (0.77). However,

failed predictions more frequently had probabilities >0.95 than the out-of-distribution samples (Figure 6b).

For both tests involving the within- and between-dataset predictions, to detect the failed predictions, we set conservative threshold values for the prediction probabilities and marked samples below the threshold as potential misclassification. When the threshold value was set to 0.95, 92% of successful predictions were retained while 76% of failures and 75% of out-of-distribution samples were correctly detected as misclassifications (Figure 6a,b).

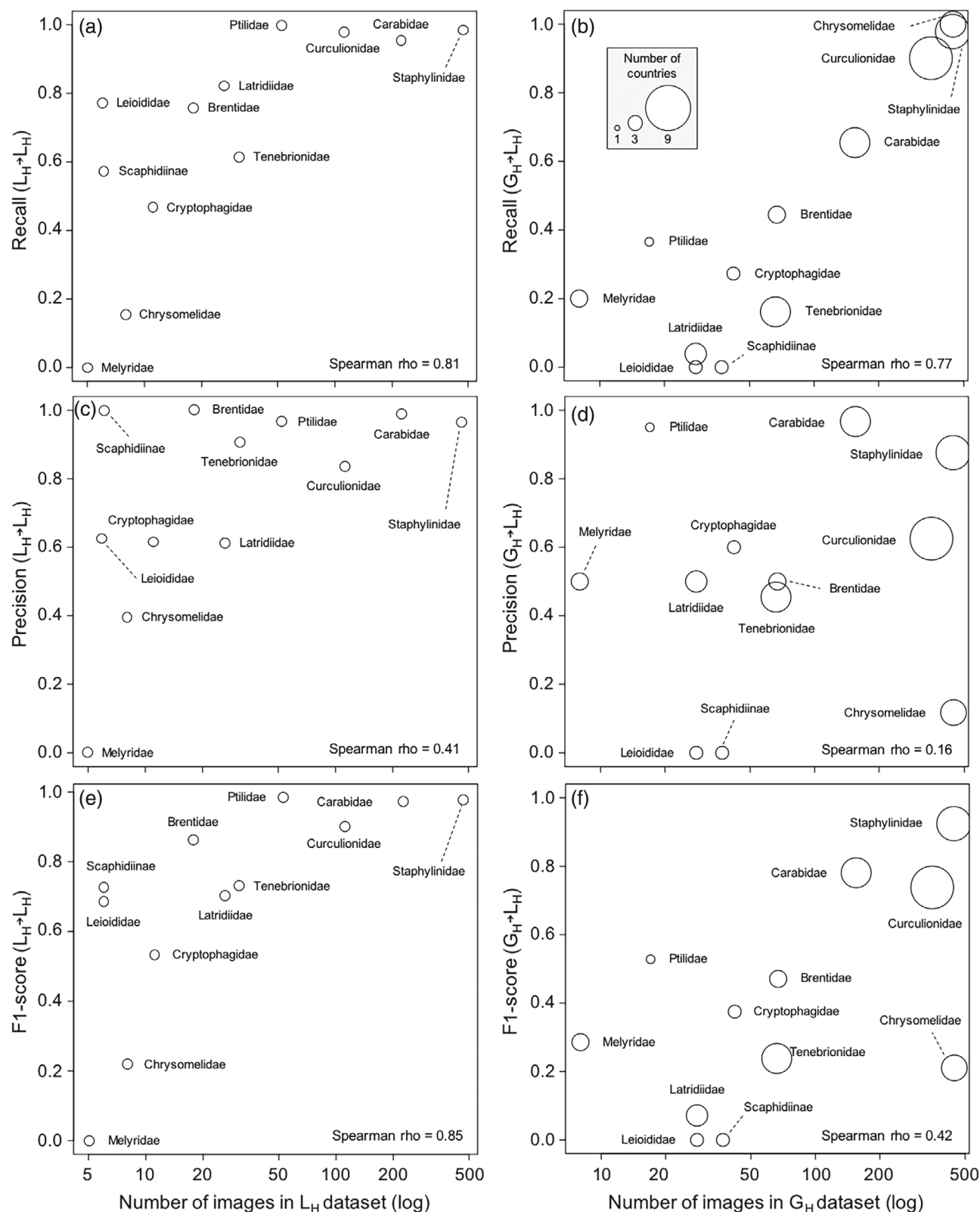


FIGURE 5 Effect of the increasing number of images on recall rates (panels a and b), multiclass precision (panels c and d) and F1-scores (panels e and f). We used 400 randomly selected *local high quality* (L_H) images for training and predicting the class (family label) of L_H images (within-dataset classification) (left panels), and 800 randomly selected *global high quality* (G_H) images for training and predicting the class of L_H images (right panels). Note that x-axes representing the number of images are on a logarithmic scale. Circle sizes represent the number of countries where samples of a given family were collected from (as a proxy of intra-family morphological variation).

DISCUSSION

This work adds to the growing number of studies demonstrating the power of CNNs in image-based taxonomic classification. Specifically, we tested the possibility of classifying specimens from bulk samples of beetles, whereby unknown local samples were classified using a model trained on similarly photographed bulk samples from a global set. We envision that mixed trap samples in future will be routinely photographed with high-resolution cameras, producing huge numbers of valuable images, but unlike most existing studies that use pinned or cardboard-glued specimens, these images present specimens in diverse angles, habitus, magnification and lighting (Schneider et al., 2022; Wühl et al., 2022). We show that these images provide sufficient information for specimens to be identified as members of particular families of Coleoptera. This finding is of special relevance in the context of large-scale biodiversity surveys where higher-rank taxonomic classification arises as a mandatory first-step prior to more refined classification by expert taxonomists (Karlsson et al., 2020). Within a local dataset, classification accuracy regularly reached 95% or more, which is similar to findings from other studies using more standardized photographs from museum collections (e.g., ~92% and 96% for Diptera and Coleoptera, respectively; Valan et al., 2019). We also confirm that classification performance depends on the number of images used for training (Figures 2–5), as widely seen in image recognition applications generally (Donahue et al., 2013) and in insect classification in particular (e.g., >90% recall rates were obtained for taxa with >50 images; Valan et al., 2019, 2021). We find that the prediction accuracy generally does not increase further after about 200 images in each of the three datasets used here. However, the degree of accuracy is greatly affected by the image quality and the complexity of the dataset: both the L_L (low image quality) and in

particular the G_H (high complexity) datasets show comparatively lower accuracy of predictions if trained on themselves.

Utility of global databases for classifying local faunas

The critical question in this study is about the success of transfer learning in a situation where the source and target data are from different faunas. We here used the challenging case of the soil fauna of a Mediterranean island as the domain target for images trained on a set of mixed trap samples from altogether 11 tropical forest sites across the globe (the G_H set), which presumably do not share any species or genera. However, most local bulk samples, even from such disparate ecosystems, share a similar set of taxa at the family level, especially for a few species-rich families which are found in similar relative proportions in most samples. The complexity of the data may allow the CNN model trained on this broad set to capture general family traits of the global fauna and thus make it suitable for a greater range of classification tasks at local level. However, this broad scope comes at a certain price, as the source accuracy is fairly low (comparing the $G_H \rightarrow G_H$ with the $L_H \rightarrow L_H$, Figure 2), but if we accept the slightly lower accuracy, our study confirms the possibility of classifying local samples against this global set. We conclude that it is not strictly necessary to create local reference databases for training, when targeting higher taxonomic levels. This finding opens the way for local biodiversity assessment studies around the globe using a universal training set. Global databases have the additional advantage of offering high numbers of images per taxon, which is more difficult to obtain locally, although it is critical for increasing the performance of the CNN-based classification (Figure 2; Donahue et al., 2013; Valan et al., 2019, 2021).

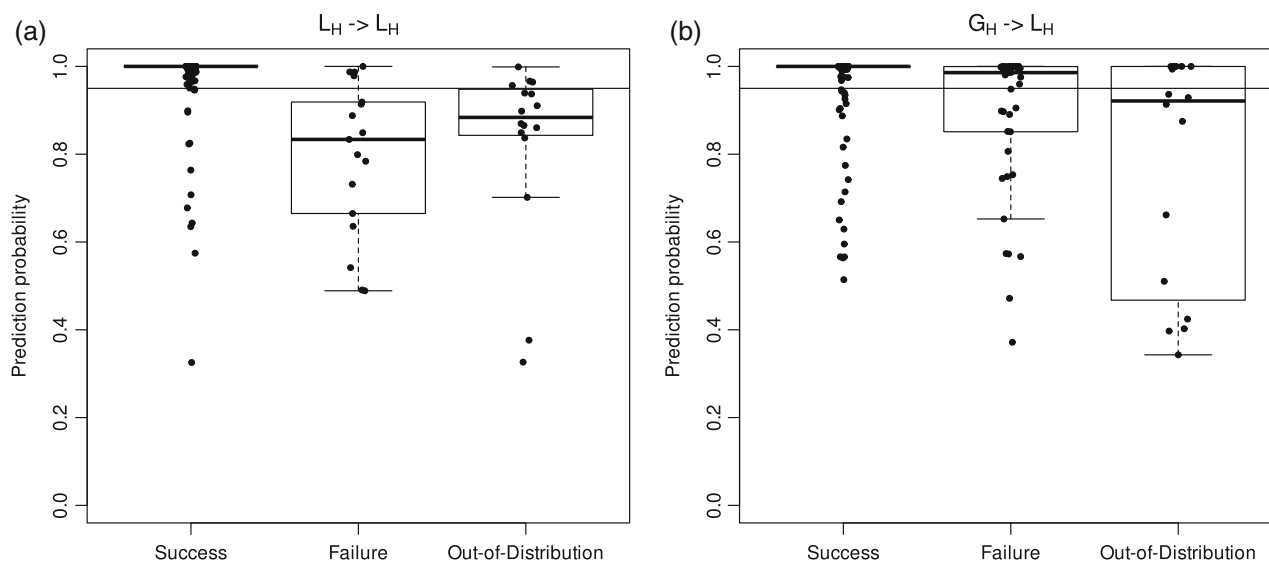


FIGURE 6 Prediction probabilities for the successful, failed and out-of-distribution predictions at a 0.95 threshold (horizontal line). (a) Intra-dataset predictions of L_H images using 400 randomly selected images for training. (b) Predictions of L_H images using 800 G_H images for training.

Despite the high prediction accuracy of the dataset as a whole, some taxa may show consistently lower classification performance. The primary factor affecting recall and precision is the number of images per family. The required quantity was available only for the largest families (which were also available for the greatest number of countries globally, as a measure of complexity of the training set). The fact that most taxa accumulate fewer single-specimen photographs as a result of rarity and low abundances, may be artificially addressed by data augmentation techniques, which have been successfully applied to specimen identification tasks (Klasen et al., 2022). However, a few taxa, including the widely sampled Chrysomelidae, showed low F1-scores even with a large number of images (Figure 5b,d,f). This example is particularly striking because of the high recall rate, but low precision, that is, although most specimens of Chrysomelidae in the sample are identified with a high prediction probability, the model misclassifies a lot of them and incorrectly assigns specimens of other families to them. The Chrysomelidae behaves poorly against the local L_H model, but this is commensurate with a low representation of images (Figure 5a,c,e). The finding may suggest a negative impact of within-family morphological disparity on classification precision, possibly only present in the wider G_H dataset. Interestingly, Chrysomelidae also showed low classification performance in the study of Valan et al. (2019). The family is composed of morphologically rather distinct subfamilies, and an increased number of images may help to unveil the subclasses generating low performance models.

Lessons from combining DANN with differing databases

We show that photographs taken from similar imaging setups (G_H and L_H) are readily used for between-region image classifications although images taken by a conventional stereoscope (L_L) exhibited a large accuracy reduction for the prediction of the local high-quality dataset. Considering the nearly identical taxonomic composition of the L_H and L_L datasets, the large accuracy reduction indicates a negative impact of the original image quality and the lack of standardization between the target–source pairs. The overall dissimilarity of L_H from G_H and L_L measured by dataset classification errors also suggest a negative effect of non-standardized imaging on prediction performance. These results are in accordance with the reduction in classification accuracy observed by other studies comparing different imaging procedures, for example, training with high-resolution museum specimens to predict field images (Knyshov et al., 2021). The application of alternative algorithms may overcome limitations resulting from the usage of highly different images taken by unstandardized imaging conditions. In the current study, we could successfully ameliorate the accuracy reductions between L_H and L_L using DANN, a method designed for improving domain adaptation (Ganin et al., 2016). However, in other combinations of datasets such as G_H and L_H , the DANN did not improve the target prediction performance. This may be due to poor hyperparameter tuning or insufficient training of the model with a complex loss function (Kouw & Loog, 2021). Nevertheless, our study would offer some evidence that DANN (or domain adaptation

techniques in general) can be considered a method of choice when a standardized image acquisition is not available.

Improvements from using alternative metrics for model performance

Although CNN-based image classification for biodiversity assessment is becoming increasingly popular, its performance is not always assessed with a broad set of performance metrics. As observed in Chrysomelidae, the reduction of performance was only detectable in the multiclass precisions and F1-scores, but not in the recalls, which revealed a specific difficulty in the classification of this group. Given the inferential power of these performance metrics, we encourage their integration in biodiversity-related applications.

Another overlooked metric is the confidence of predictions. We could detect failed predictions and potential out-of-distribution samples by setting a threshold value on the probabilities. In accordance with Hendrycks and Gimpel (2017), such misclassified or out-of-distribution samples were predicted with consistently lower prediction probabilities. Because out-of-distribution samples are common in biodiversity surveys, detection of unknown target samples based on low prediction confidence is particularly useful. A potential difficulty of this approach is that calibration of the threshold requires extra data. Conventional deep neural networks can be uncalibrated, that is, prediction probabilities do not precisely reflect prediction accuracy (Guo et al., 2017). Such uncalibrated models can make an incorrect prediction with excessively high confidence. This overconfident failure is noticeable in our analysis (Figure 6b). Therefore, additional labelled samples are required to set a robust threshold for the identification of failure and out-of-distribution samples. Methods for explicit calibration of prediction probabilities or detection of out-of-distribution samples without additional data (e.g., Hsu et al., 2020; Mukhoti et al., 2020) are being actively developed in the machine learning field, and applying those methods is a potential future direction. As DANN could remove the dataset biases caused by the imaging instruments, the purpose-specific models will expand the possibility of machine learning applications to biodiversity surveys (see Høye et al., 2021).

Building the global database for CNN-based classification

As new images become available for ever more species, the reference library for taxonomic identification is rapidly growing. Given the geographic and taxonomic distance of our reference set from tropical forests, the family category is the only meaningful level exhibiting overlap of source and target, but conceivably the methodology could be applied at lower levels, for example, genera, if more similar samples had been used. The current set of images is limited with regard to the number of families (classes) and number of images per family (intra-class variability), resulting in out-of-distribution errors and prediction errors, respectively. Both issues can be addressed with a wider

selection of images, for example, those available from the SITE-100 project (Bian et al. 2022) taken with similar equipment. Based on our results, any future image collection should consider the need for standardization, including that imaging should use the same aspect, for example, dorsal view for Coleoptera (also see Hansen et al., 2020), uniform background across images (preferably a light colour without texture), clear separation of specimens in the photographs, and similar optical equipment and magnification. The exact parameters remain to be explored within and across studies, but standardization of imaging is critical to transferability when rolling out large-scale efforts for image-based classification in biodiversity studies. As part of this effort, image segmentation should be improved and automated (Schneider et al., 2022; Schwartz & Alfaro, 2021), to increase our capability for rapidly generating ‘clean’ and individual-based image databases extracted from bulk samples. A potential bottleneck is the need to expand the training set gradually, which generally requires recomputation of the model when new classes are added, although recent updated methods may simplify this process (Hadsell et al., 2020). A second issue affecting the accuracy of predictions is the ‘category bias’ from inconsistent categorisation and labelling of the training set itself. In the current study, images in the training set were classified from the images by recognizing the overall gestalt of a family. These family labels were straightforward for most groups, but identification of some beetle families may be compromised due to images that obscured appendages or other key traits, especially in small-bodied Leiodidae, Latridiidae or Cryptophagidae, which may have contributed to the prediction errors seen in these families (Table S4). Thus, corrections to the family labels in the database may be required, possibly by DNA barcoding and phylogenetic placement methods that confirm the family membership. Likewise, combining image acquisition for biodiversity assessment with metabarcoding could be instrumental for validating or improving genetic-based inferences (Yang et al., 2022) or estimating biomass and abundance (e.g., Høye et al., 2021; Schneider et al., 2022). Metabarcoding studies often lose morphological information of specimens, but imaging could be accommodated as a routine step before the DNA extraction of bulk insect samples.

CONCLUSIONS

To our knowledge, this is the first attempt of domain adaptation for taxonomic classification of an entirely unknown dataset, as a key element of using image-based identification in biodiversity studies at the global scale. We show that the approach is highly feasible, but needs careful consideration of the imaging procedure, the algorithmic approach and the choice of training sets. We envisage that an increasingly complete set of images, covering the diversity of major taxonomic groups, will become available as a global database in future, against which samples from any ecosystem and biogeographic region can be classified at a certain hierarchical level (e.g., families of beetles). In our approach, we lack the close alignment of the feature space in source and target that would guarantee high transferability, albeit at the expense of lower generalization capability when encountering unknown samples.

Further studies are required to assess the trade-offs of broadening the source domain and to establish best practice for the specific research question at hand. Once a stable expanded image database has been created, it can be used for wider applications in biodiversity research and monitoring, potentially building a global model applicable to any sampling site and possibly used while still in the field.

ACKNOWLEDGMENTS

This work was supported by the iBioGen project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810729. We are grateful to Richard Turney and Thomas J. Creedy (Natural History Museum, London) for advice and support during bulk-sample imaging, and Takashi Imai (Shiga University) for helpful advice on deep learning methods. We also thank three anonymous referees for their constructive and valuable comments on an earlier version of the manuscript. We would like to extend our gratitude to Andreas Dimitriou for help during sample imaging, and Konstantinos Ntatsopoulos for support in the taxonomy of Cyprus beetles. Víctor Noguerales was supported by a postdoctoral contract under the iBioGen project and a “Juan de la Cierva-Formación” postdoctoral fellowship (grant: FJC2018-035611-I) funded by MCIN/AEI/10.13039/501100011033. Tomochika Fujisawa was supported by JSPS KAKENHI (grant number: 20K06824).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The code to reproduce the analysis is available for download from GitHub https://github.com/tfujisawa/bulk_image_classification. The image databases are available in the Dryad Digital Repository (<https://doi.org/10.5061/dryad.05qfttf4f>) and Zenodo (https://zenodo.org/record/5823545#.Y5_bpxVBwuV).

ORCID

Tomochika Fujisawa  <https://orcid.org/0000-0002-4611-3727>

Víctor Noguerales  <https://orcid.org/0000-0003-3185-778X>

Emmanouil Meramveliotakis  <https://orcid.org/0000-0002-6399-575X>

Anna Papadopoulou  <https://orcid.org/0000-0002-4656-4894>

Alfried P. Vogler  <https://orcid.org/0000-0002-2462-3718>

REFERENCES

- Årje, J., Melvad, C., Jeppesen, M.R., Madsen, S.A., Raitoharju, J., Rasmussen, M.S. et al. (2020) Automatic image-based identification and biomass estimation of invertebrates. *Methods in Ecology and Evolution*, 11(8), 922–931. Available from: <https://doi.org/10.1111/2041-210X.13428>
- Arribas, P., Andújar, C., Hopkins, K., Shepherd, M. & Vogler, A.P. (2016) Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution*, 7(9), 1071–1081. Available from: <https://doi.org/10.1111/2041-210X.12557>
- Basset, Y., Cizek, L., Cuénoud, P., Didham, R.K., Guilhaumon, F., Missa, O. et al. (2012) Arthropod diversity in a tropical forest. *Science*,

- 338(6113), 1481–1484. Available from: <https://doi.org/10.1126/science.1226727>
- Bian, X., Garner B.H., Liu, H. & Vogler A.P. (2022) The SITE-100 project: Site-based biodiversity genomics for species discovery, community ecology, and a global tree-of-life. *Frontiers in Ecology and Evolution*, 10, 787560. <https://doi.org/10.3389/fevo.2022.787560>
- Buschbacher, K., Ahrens, D., Espeland, M. & Steinhage, V. (2020) Image-based species identification of wild bees using convolutional neural networks. *Ecological Informatics*, 55, 101017. Available from: <https://doi.org/10.1016/j.ecoinf.2019.101017>
- Caruso, T., Schaefer, I., Monson, F. & Keith, A.M. (2019) Oribatid mites show how climate and latitudinal gradients in organic matter can drive large-scale biodiversity patterns of soil communities. *Journal of Biogeography*, 46(3), 611–620. Available from: <https://doi.org/10.1111/jbi.13501>
- Costello, M.J., May, R.M. & Stork, N.E. (2013) Can we name Earth's species before they go extinct? *Science*, 339(6118), 413–416. Available from: <https://doi.org/10.1126/science.1230318>
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. et al. (2013) DeCAF: a deep convolutional activation feature for generic visual recognition. *ICML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning*, 32, 1–647–1–655. Available from: <https://doi.org/10.5555/3044805.3044879>
- Escribano, N., Oscoz, J., Galicia, D., Cancellario, T., Durçan, C., Navarro, P. et al. (2018) Freshwater macroinvertebrate samples from a water quality monitoring network in the Iberian Peninsula. *SciData*, 5, 180108. Available from: <https://doi.org/10.1038/sdata.2018.108>
- Farahani, A., Voghoei, S., Rasheed, K. & Arabnia, H.R. (2020) A brief review of domain adaptation. *arXiv 2010.03978v1*. <https://doi.org/10.48550/arXiv.2010.03978>
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F. et al. (2016) Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59), 1–35. Available from: <http://jmlr.org/papers/v17/15-239.html>
- Guan, H. & Liu, M. (2021) Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 1–1, 1173–1185. Available from: <https://doi.org/10.1109/tbme.2021.3117407>
- Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q. (2017) On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1321–1330. Available from: <https://proceedings.mlr.press/v70/guo17a.html>
- Hadsell, R., Rao, D., Rusu, A.A. & Pascanu, R. (2020) Embracing change: continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12), 1028–1040. Available from: <https://doi.org/10.1016/j.tics.2020.09.004>
- Hansen, O.L.P., Svenning, J.C., Olsen, K., Dupont, S., Garner, B.H., Iosifidis, A. et al. (2020) Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and Evolution*, 10(2), 737–747. Available from: <https://doi.org/10.1002/ece3.5921>
- Hendrycks, D. & Gimpel, K. (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the 5th international conference on learning representations, ICLR 2017*. Available from: <https://arxiv.org/abs/1610.02136>
- Høye, T.T., Årje, J., Bjerre, K., Hansen, O.L.P., Iosifidis, A., Leese, F. et al. (2021) Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences of the United States of America*, 118(2), e2002545117. Available from: <https://doi.org/10.1073/pnas.2002545117>
- Hsu, Y.C., Shen, Y., Jin, H. & Kira, Z. (2020) Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10948–10957. Available from: <https://doi.org/10.1109/CVPR42600.2020.01096>
- Hudson, L.N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B.W. et al. (2015) INSELECT: automating the digitization of natural history collections. *PLoS One*, 10(11), e0143402. Available from: <https://doi.org/10.1371/journal.pone.0143402>
- Karlsson, D., Forshage, M., Holston, K. & Ronquist, F. (2020) The data of the Swedish malaise trap project, a countrywide inventory of Sweden's insect fauna. *Biodiversity Data Journal*, 8, e56586.
- Klasen, M., Ahrens, D., Eberle, J. & Steinhage, V. (2022) Image-based automated species identification: can virtual data augmentation overcome problems of insufficient sampling? *Systematic Biology*, 71(2), 320–333. Available from: <https://doi.org/10.1093/sysbio/syab048>
- Knyshov, A., Hoang, S. & Weirauch, C. (2021) Pretrained convolutional neural networks perform well in a challenging test case: identification of plant bugs (Hemiptera: Miridae) using a small number of training images. *Insect Systematics and Diversity*, 5(2) 3, 1–10. Available from: <https://doi.org/10.1093/isd/ixab004>
- Kouw, W.M. & Loog, M. (2021) A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3), 766–785. Available from: <https://doi.org/10.1109/TPAMI.2019.2945942>
- LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, 521(7553), 436–444. Available from: <https://doi.org/10.1038/nature14539>
- Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H.S. & Dokania, P. K. (2020) Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33, 15288–15299. Available from: <https://proceedings.neurips.cc/paper/2020/hash/aeb7b30ef1d024a76f21a1d40e30c302-Abstract.html>
- Noguerales, V., Meramveliotakis, E., Castro-Insua, A., Andújar, C., Arribas, P., Creedy, T.J. et al. (2021) Community metabarcoding reveals the relative role of environmental filtering and spatial processes in metacommunity dynamics of soil microarthropods across a mosaic of montane forests. *Molecular Ecology* in press. Available from: <https://doi.org/10.1111/mec.16275>
- Novotny, V., Miller, S.E., Hulcr, J., Drew, R.A.I., Basset, Y., Janda, M. et al. (2007) Low beta diversity of herbivorous insects in tropical forests. *Nature*, 448(7154), 692–695. Available from: <https://doi.org/10.1038/nature06021>
- Pan, S.J. & Yang, Q.Y. (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359. Available from: <https://doi.org/10.1109/TKDE.2009.191>
- Popkov, A., Konstantinov, F., Neimorovets, V. & Solodovnikov, A. (2022) Machine learning for expert-level image-based identification of very similar species in the hyperdiverse plant bug family Miridae (Hemiptera: Heteroptera). *Systematic Entomology*, 47(3), 487–503. Available from: <https://doi.org/10.1111/syen.12543>
- R Core Team. (2021) R: a language and environment for statistical computing. Available from: <https://www.r-project.org/>
- Raitoharju, J., Riabchenko, E., Ahmad, I., Iosifidis, A., Gabbouj, M., Kiranyaz, S. et al. (2018) Benchmark database for fine-grained image classification of benthic macroinvertebrates. *Image and Vision Computing*, 78, 73–83. Available from: <https://doi.org/10.1016/j.imavis.2018.06.005>
- Razavian, A.S., Azizpour, H., Sullivan, J. & Carlsson, S. (2014) CNN features off-the-shelf: an astounding baseline for recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 806–813. Available from: https://www.cv-foundation.org/openaccess/content_cvpr_workshops_2014/W15/html/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.html
- Romero, I.C., Kong, S., Fowlkes, C.C., Jaramillo, C., Urban, M.A., Oboh-Ikuenobe, F. et al. (2020) Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 117(45), 28496–28505. Available from: <https://doi.org/10.1073/pnas.2007324117>
- Schneider, S., Taylor, G.W., Kremer, S.C., Burgess, P., McGroarty, J., Mitsui, K. et al. (2022) Bulk arthropod abundance, biomass, and

- diversity estimation using deep learning for computer vision. *Methods in Ecology and Evolution*, 13(2), 346–357. Available from: <https://doi.org/10.1111/2041-210X.13769>
- Schwartz, S.T. & Alfaro, M.E. (2021) Sashimi: a toolkit for facilitating high-throughput organismal image segmentation using deep learning. *Methods in Ecology and Evolution*, 12(12), 2341–2354. Available from: <https://doi.org/10.1111/2041-210X.13712>
- Stork, N.E. & Grimbacher, P.S. (2006) Beetle assemblages from an Australian tropical rainforest show that the canopy and the ground data contribute equally to biodiversity. *Proceedings of the Royal Society B*, 273(1596), 1969–1975. Available from: <https://doi.org/10.1098/rspb.2006.3521>
- Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., Vercauteren, K.C., Snow, N.P. et al. (2019) Machine learning to classify animal species in camera trap images: applications in ecology. *Methods in Ecology and Evolution*, 10(4), 585–590. Available from: <https://doi.org/10.1111/2041-210X.13120>
- Tommasi, T., Patricia, N., Caputo, B. & Tuytelaars, T. (2017) A deeper look at dataset bias. In: Csurka, G. (Ed.) *Domain adaptation in computer vision applications. Advances in computer vision and pattern recognition*. Cham, Switzerland: Springer, pp. 37–55. Available from: https://doi.org/10.1007/978-3-319-58347-1_2
- Torralba, A. & Efros, A.A. (2011) Unbiased look at dataset bias, *CVPR* 2011, 1521–1528, Available from: <https://doi.org/10.1109/CVPR.2011.5995347>
- Valan, M., Makonyi, K., Maki, A., Vondráček, D. & Ronquist, F. (2019) Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Systematic Biology*, 68(6), 876–895. Available from: <https://doi.org/10.1093/sysbio/syz014>
- Valan, M., Vondráček, D. & Ronquist, F. (2021) Awakening a taxonomist's third eye: exploring the utility of computer vision and deep learning in insect systematics. *Systematic Entomology*, 46(4), 757–766. Available from: <https://doi.org/10.1111/syen.12492>
- Wührl, L., Pylatiuk, C., Giersch, M., Lapp, F., von Rintelen, T., Balke, M. et al. (2022) DiversityScanner: robotic handling discovery of small invertebrates with machine learning methods. *Molecular Ecology Resources*, 22(4), 1626–1638. Available from: <https://doi.org/10.1111/1755-0998.13567>
- Yang, B., Zhang, Z., Yang, C.-Q., Wang, Y., Orr, M.C., Wang, H. et al. (2022) Identification of species by combining molecular and morphological data using convolutional neural networks. *Systematic Biology*, 71(3), 609–705. Available from: <https://doi.org/10.1093/sysbio/syab076>

SUPPORTING INFORMATION

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

Figure S1. Detailed scheme of architectures of the convolutional neural network (CNN) and domain adversarial neural networks (DANN) used in this study.

Figure S2. Loss and accuracy development during training of CNN models.

Figure S3. Loss and accuracy development during training of DANN models for three pairs of data sets. Test loss and test accuracy were measured on the target data.

Figure S4. Effect of the number of images on prediction accuracy for the *Local High Quality* (L_H) and *Global High Quality* (G_H) images.

Figure S5. Effects of the number of images on prediction accuracy for the *Local Low Quality* (L_L) and *Global High Quality* (G_H) images.

Figure S6. Exemplar images of incorrect classification.

Table S1. Description of the machine learning terminology used in this study.

Table S2. Number of images per taxa and dataset used in the present study.

Table S3. Details of sampling sites for the *Global High Quality* (G_H) dataset.

Table S4. A confusion matrix for the prediction of *Local High Quality* (L_H) images based on a training set of 400 randomly selected images from the same dataset (L_H).

Table S5. A confusion matrix for the prediction of *Local High Quality* (L_H) images based on a training set of 800 randomly selected *Global High Quality* (G_H) images.

How to cite this article: Fujisawa, T., Nogueras, V., Meramveliotakis, E., Papadopoulou, A. & Vogler, A.P. (2023) Image-based taxonomic classification of bulk insect biodiversity samples using deep learning and domain adaptation. *Systematic Entomology*, 48(3), 387–401. Available from: <https://doi.org/10.1111/syen.12583>