

Taxonomic Classification of Ants (Formicidae) from Images using Deep Learning

MARIJN J. A. BOER¹ AND RUTGER A. VOS^{1,*}

¹ *Endless Forms, Naturalis Biodiversity Center, Leiden, 2333 BA, Netherlands*

**rutger.vos@naturalis.nl*

ABSTRACT

The well-documented, species-rich, and diverse group of ants (Formicidae) are important ecological bioindicators for species richness, ecosystem health, and biodiversity, but ant species identification is complex and requires specific knowledge. In the past few years, insect identification from images has seen increasing interest and success, with processing speed improving and costs lowering. Here we propose deep learning (in the form of a convolutional neural network (CNN)) to classify ants at species level using AntWeb images. We used an Inception-ResNet-V2-based CNN to classify ant images, and three shot types with 10,204 images for 97 species, in addition to a multi-view approach, for training and testing the CNN while also testing a worker-only set and an AntWeb protocol-deviant test set. Top 1 accuracy reached 62% - 81%, top 3 accuracy 80% - 92%, and genus accuracy 79% - 95% on species classification for different shot type approaches. The head shot type outperformed other shot type approaches. Genus accuracy was broadly similar to top 3 accuracy. Removing reproductives from the test data improved accuracy only slightly. Accuracy on AntWeb protocol-deviant data was very low. In addition, we make recommendations for future work concerning image threshold, distribution, and quality, multi-view approaches, metadata, and on protocols; potentially leading to higher accuracy with less computational effort.

¹⁸ *Key words:* Deep Learning, Convolutional Neural Network, Image Classification,
¹⁹ Formicidae, AntWeb, multi-view

²⁰

²¹ The family of ants (Formicidae) is a large and diverse group within the insect order,
²² occasionally exceeding other insect groups in local diversity by far. Representing the bulk
²³ of global biodiversity (Mora et al. 2011), ants are globally found (except on Antarctica)
²⁴ and play important roles in a lot of ecosystems (Hölldobler et al. 1990). As ants are found
²⁵ to be good bioindicators, ecological and biodiversity data on them may be used to assess
²⁶ the state of ecosystems (Andersen 1997; Andersen et al. 2002), which is important for
²⁷ species conservation. Furthermore, insects are good surrogates for predicting species
²⁸ richness patterns in vertebrates because of their significant biomass (Andersen 1997;
²⁹ Moritz et al. 2001), even while using the morphospecies concept (Oliver et al. 1996; Pik
³⁰ et al. 1999). To understand the ecological role and biological diversity of ants, it is
³¹ important to comprehend their morphology, and delimit and discriminate among species.
³² Even working with morphospecies, a species concept is still required for identification to
³³ reach a level of precision sufficient to answer a research question. This is what is called
³⁴ Taxonomic Sufficiency (Ellis 1985), which must be at a certain balance or level for a
³⁵ research goal (Groc et al. 2010). Therefore, it is important to get a good understanding of
³⁶ ant taxonomy, but many difficulties arise with the complicated identification of ants to
³⁷ species level or to taxonomic sufficiency.

³⁸

Ant taxonomy

³⁹ Classifying and identifying ant species is complex work and requires specific
⁴⁰ knowledge. While there is extensive work on this (e.g. Bolton (1994), Fisher et al. (2007),
⁴¹ and Fisher et al. (2016)), it is still in many instances reserved to specialists. To identify ant
⁴² species, taxonomists use distinct characters (e.g. antennae, hairs, carinae, thorax shape,
⁴³ body shininess) that differ between subfamilies, genera, and species. However, the detailed

knowledge on morphological characters can sometimes make species identification difficult. Some ant species appear to be sibling species or very cryptic, and different castes complicate things further. However, with a long history on myrmecological research, ants are one of the best documented groups of insects and in recent years ant systematics have seen substantial progress (Ward 2007).

49 *Computer vision*

In an effort to improve taxonomic identification, insect identification from images has been a subject of computer vision research in the past few years. As some early papers have shown (D. E. Guyer et al. 1986; Edwards et al. 1995; PJD Weeks et al. 1997; PJ Weeks et al. 1999; Gaston et al. 2004), a promising start has been made on automated insect identification, but there is still a long road to reaching human accuracy. Systems like a Bayes classifier (D. E. Guyer et al. 1986) or DAISY ((PJ Weeks et al. 1999) mostly utilized structures, morphometrics, and outlines. Together with conventional classifying methods (such as a principal component analysis (PCA) (P Weeks et al. 1997)) images data could be classified. Other, slightly more complex systems use simple forms of machine learning (ML) (Kang et al. 2012), such as a support vector machine (SVM) ((Yang et al. 2015) or K -nearest neighbors (Watson et al. 2004). An identification system for insects at the order level (including ants within the order of Hymenoptera) designed by Wang et al. (2012b), used seven geometrical features (e.g. body width) and reached 97% accuracy. Unfortunately, there are no classification studies that include ants, outside of the work of Wang et al. (2012b) on insect order level, but for other insect groups, promising results have been reported. Butterflies families (Lepidoptera) have been identified using shape, color and texture features, exploiting the so-called CBIR algorithm (Wang et al. 2012a). Insect identification to species level is harder, as some studies have shown. Javanese butterflies (Lepidoptera: Nymphalidae, Pieridae, Papilionidae, and Riodinidae) could be discriminated using the BGR-SURF algorithm with 77% accuracy (Vetter 2016). Honey

bees (Hymenoptera: Apidae) could be classified with good results (>90%), using wing morphometrics with multivariate statistics (Francoy et al. 2008). Gerard et al. (2015) could discriminate haploid and diploid bumblebees (Hymenoptera: Apidae) based on differences in wing shape (e.g. wing venation patterns) with great success (95%). Seven owlfly species (Neuroptera: Ascalaphidae) were classified using an SVM on wing outlines (99%) (Yang et al. 2015). Five wasp species (Hymenoptera: Ichneumonidae) could be classified using PCA on wing venation data (94%) (P Weeks et al. 1997). Wen et al. (2012) classified eight insect species (Tephritidae and Tortricidae) using 54 global morphological features with 86.6% accuracy. And Kang et al. (2012) fed wing morphometrics for seven butterfly species (Lepidoptera: Nymphalidae and Papilionidae) in a simple neural network to classify, resulting in >86% accuracy. However, a significant disadvantage in these systems is the need for metric morphological features exploitation, which still require human expertise, supervision, and input.

Deep learning

Deep learning (DL) may therefore be a promising taxonomic identification tool, as it does not require human supervision. DL allows a machine to learn representations of features by itself, instead of conventional methods where features need manual introduction to the machine (Bengio et al. 2012; LeCun et al. 2015). In the past few years, DL has attracted attention in research and its methods and algorithms have greatly improved, which is why its success will likely grow in the future (LeCun et al. 2015). A successful DL algorithm is the convolutional neural network (CNN), mostly used for image classification and preferably trained using GPUs. These computationally-intensive networks are designed to process (convolve) 2D data (images), using typical neural layers as convolutional and pooling layers (Krizhevsky et al. 2012; LeCun et al. 2015) and can even work with multi-view approaches (Zhao et al. 2017). A simple eight layer deep CNN has strongly outperformed conventional algorithms that needed introduced features (Held

96 et al. 2015). It is also common practice that deep neural networks outperform shallow
97 neural networks (Chatfield et al. 2014). In recent years, CNN technology has advanced
98 greatly (LeCun et al. 2015; Mishkin et al. 2017; Wäldchen et al. 2018), and many
99 biological relevant studies have shown promising results (as can be read in the next
100 Section: Related deep learning studies).

101 *Related deep learning studies* CNNs have been used in plant identification (Lee
102 et al. 2015; Lee et al. 2016; Dyrmann et al. 2016; Barré et al. 2017; Sun et al. 2017), plant
103 disease detection (Mohanty et al. 2016) and identification of underwater fish images (Qin
104 et al. 2016), all with high accuracy (71% – 99%). Applied examples with high accuracy
105 include classification of different qualities of wood for industrial purposes (79%) (Affonso
106 et al. 2017), identifying mercury stained plant specimens from non-stained (90%)
107 (Schuettpelz et al. 2017), and identification of specimens using multiple herbariums (70% –
108 80%) (Carranza-Rojas et al. 2017). Especially studies like the last two are important for
109 natural history collections, because such applications can benefit research, speed up
110 identification and lower costs.

111 *Contributions*

112 Here, we explore an alternative approach to taxonomic identification of ants based
113 on computer vision and deep learning, using images from AntWeb (AntWeb.org 2017[a]).
114 AntWeb is the world's largest and leading online open database for ecological, taxonomic,
115 and natural history information on ants. AntWeb keeps records and high quality images of
116 specimens from all over the world, usually maintained by expert taxonomist curators. Ant
117 mounting and photographing of specimens usually follows the AntWeb protocol
118 (AntWeb.org 2018), which specifies standards for a dorsal, head and profile view.
119 Considering that automating identification could greatly speed up taxonomic work and
120 improve identification accuracy (PJD Weeks et al. 1997; Gaston et al. 2004), this work

121 could assist in solving collection impediments. In this research, we make a start with
122 automatic classification of ant species using images and present which shot type will
123 classify ants the best, together with a multi-view approach. In the end we will discuss the
124 results from different data sets and write recommendations for future work in an effort to
125 improve taxonomic work and increase classification accuracy.

126 MATERIALS AND METHODS

127 First presented are the data sets, the process involving quality of data and, creating
128 test sets. We used different shot types to find which type classifies best. In a different
129 approach the three shot types are combined to one image for multi-view training. More test
130 data for all shot types is a worker-only set and an AntWeb protocol-deviant set. Secondly,
131 image augmentation is described and explained. Thirdly, the proposed model with its
132 architectural decisions is discussed, and lastly the model related preprocessing actions.

133 *Data material*

134 We collected all images and metadata from AntWeb (AntWeb.org 2017[a]), where
135 images follow specific protocols for mounting and photographing with dorsal, head and
136 profile views(AntWeb.org 2018). The intention was to work with 100 species, but the list
137 was truncated at the 97 most imaged. This ensured the data included all species with 68 or
138 more images, leaving out all species with 67 images or fewer. On May 15, 2018, catalog
139 number, genus and species name, shot type, and image for imaged specimens of the 97
140 species were harvested from AntWeb, through its API version 2 (AntWeb.org 2017[b]).
141 This first data set with a total of 3,437 specimens and 10,211 images is here referred to as
142 *top97species_Qmed_def*. The distribution of images per species for the dorsal shot type
143 (3,405 images), head (3,385) and profile (3,421) can be seen in Figure 1 on page 27 and
144 Table 1 on page 34. We partitioned the images randomly in non-overlapping sets:
145 approximately 70%, 20%, and 10% for training, validation, and testing, respectively (see

¹⁴⁶ Table 1 on page 34). The 70%-20%-10% was used in every consecutive dataset involving
¹⁴⁷ training. We downloaded images in medium quality, accountable for 233 pixels in width and
¹⁴⁸ ranging from 59 pixels to 428 pixels in height (for sample images see Figure 2 on page 28).

¹⁴⁹ *Cleaning the data* This initial data set still contained specimens that miss a gaster
¹⁵⁰ and/or head or are close ups of body parts (e.g. thorax, gaster, or mandibles). A small
¹⁵¹ group of other specimens showed damage by fungi or were affected by glue, dirt or other
¹⁵² substances. These images were removed from the dataset, as these images are not
¹⁵³ representing complete ant specimens and could affect the accuracy of the model. A total of
¹⁵⁴ 94 images (46 specimens) were omitted from training, validation and testing (dorsal: 43,
¹⁵⁵ head: 7, profile: 44), resulting in 10,117 images for 3,407 specimens for a new dataset
¹⁵⁶ named *top97species_Qmed_def_clean*. Most of the images of detached heads could still be
¹⁵⁷ used, as the heads were glued on pinned paper points and looked just like non-detached
¹⁵⁸ head images.

¹⁵⁹ *Multi-view data set* In order to create a multi-view dataset we only included
¹⁶⁰ specimens in *top97species_Qmed_def_clean* with all three shot types. A total of 95
¹⁶¹ specimens (151 images) had two or fewer shot types and, thus could not be used. This list
¹⁶² was combined with the bad specimen list for a total of 115 specimens (as there was some
¹⁶³ overlap with the one/two shot specimens and bad specimens). We removed these 115
¹⁶⁴ specimens from the initial dataset so 3,322 specimens remained, all with three images per
¹⁶⁵ specimen per shot type, in a dataset named *top97species_Qmed_def_clean_multi* (see Table
¹⁶⁶ 1 on page 34). The most imaged *Camponotus maculatus* (Fabricius, 1782) had 223
¹⁶⁷ three-shot specimens and the least imaged species *Camponotus christi* (Forel, 1886) only
¹⁶⁸ 18. Before stitching, we scaled all images to the same width, using the width of the widest
¹⁶⁹ image. If after scaling an image had fewer pixels in height than the largest image, black
¹⁷⁰ pixels were added to the bottom of this image to complement the height of the largest
¹⁷¹ image (example in Figure 3 on page 29). We did not consider the black pixels as a problem

for classification, because almost all stitched images had black pixel padding. The model will therefore learn that these black pixels are not representing discriminating features between species. Now, the images were combined in a horizontally stacked dorsal-head-profile image, followed by normalizing pixel values to $[-1, 1]$ and resizing width and height to 299×299 pixels.

Worker only test set We labeled all specimens with their correct caste manually, as AntWebs API version 2 did not support the use of castes (support for this will be in version 3 (AntWeb.org 2017[c])). We considered alate, dealate and ergatoid queens, (ergatoid) males and intercastes as non-workers (i.e. reproductives), with no intercastes in the data set. Over 80% of *top97species_Qmed_def_clean* appeared to be workers (Figure1b on page 27). Consequently, 651 specimens (1,831 images) were marked as reproductives, with potential exclusion from a test set copy of *top97species_Qmed_def_clean*. A total of 63, 52 and 58 images, for dorsal, head, profile respectively, were removed from this copy to create a test set named *top97species_Qmed_def_clean_wtest*. The number of images in *top97species_Qmed_def_clean_wtest* set are 264, 279 and 278 for dorsal, head and profile, respectively (see Table 1 on page 34). Unfortunately, for a few species all test images were from reproductive specimens, resulting in no test images for that species. The dorsal set had five species with no test data, head only one and profile three.

St. Eustatius 2015 collection In a 2015 expedition to St. Eustatius, researchers of Naturalis Biodiversity Center collected an extensive amount of flora and fauna (Andel et al. 2016). During this expedition, researchers also collected a considerable number of ant samples, now stored at Naturalis Biodiversity Center, in Leiden, the Netherlands. Most of these species all had one or more specimens imaged, and the majority of this collection was identified by expert ant taxonomists. From this collection, we extracted images of species shared with *top97species_Qmed_def* in a new data set we refer to as *statia2015_rmnh*. This test data set of seven species with 28 images per shot type (see Table 1 on page 34) is used

198 to assess whether the model can be applied to AntWeb protocol-deviant collections,
199 indicating if an application will be of practical use to natural history museums and
200 collections with existing image banks.

201 *Data augmentation*

202 The issue of a small data set (<1 million training images) can be tackled by using
203 image augmentation, a very common method used in DL (Krizhevsky et al. 2012). In order
204 to artificially increase the training set, we applied label-preserving image augmentation
205 randomly to training images during the forward pass in the training phase. Images were
206 randomly rotated between -20° and 20° , vertically and horizontally shifted between 0%
207 and 20% of the total height and width, horizontally sheared for maximally 20° , zoomed in
208 for maximally 20% and horizontally flipped. It did not make sense to do heavier or other
209 transformations, e.g. vertical flipping as ant images will never be upside down. With data
210 augmentation, model performance is boosted because the model becomes more robust to
211 inconsistencies in ant mounting and to within-species variation. Data augmentation can
212 decrease the error rate between training and test accuracy, and therefore reduce overfitting
213 (Wong et al. 2016). For data augmentation examples see Figure 4 on page 30.

214 *Deep learning framework and model*

215 We did all of the programming in Python, mostly utilizing the open source deep
216 learning framework Keras (Chollet 2015), with the TensorFlow framework as backend
217 (Abadi et al. 2016). We ran all experiments on a Windows 10 (64 bit) computer with a
218 3.50 GHz Intel Xeon E5-1650 v3 CPU and an Nvidia GeForce GTX Titan X (12GB). The
219 network we used was Inception-ResNet-V2 (Szegedy et al. 2016) because of its efficient
220 memory usage and computational speed. We added four top layers for this classification
221 problem to create a modified version of Inception-ResNet-V2 (Fig 5 on page 31), in order:

- 222 1. Global average pooling layer to minimize overfitting and reduce model parameters

223 (Lin et al. 2013).

224 2. Dropout layer with 50% dropout probability to minimize overfitting (Srivastava et al.

225 2014).

226 3. Fully connected layer with the ReLU function as activation (Glorot et al. 2011).

227 4. Fully connected softmax layer to average prediction scores to a distribution over 97

228 classes (Krizhevsky et al. 2012).

229 As transfer learning is found to be a favorable method during training (Yosinski

230 et al. 2014), we initialized with pre-trained weights (for inception models trained by

231 Keras-team (MIT license) using the ImageNet data set (Deng et al. 2009)). We found

232 transfer learning and fine-tuning from ImageNet to be consistently beneficial in training the

233 ant classification models (no layers were frozen) as it greatly decreased training time. To

234 update the parameters we used the Nadam optimizer (Dozat 2016), which is a modification

235 of the Adam optimizer (Kingma et al. 2014) using Nesterov momentum. Nesterov

236 momentum is usually superior to vanilla momentum (Ruder 2016), which is used in Adam.

237 We initialized Nadam with standard Keras settings (e.g. $decay = 0.004$), except one: the

238 learning rate was set to 0.001 and allowed to change if model improvement stagnated.

239 *Preprocessing*

240 Before training, we normalized pixel values to $[-1, 1]$ to meet the requirements of

241 Inception-ResNet-V2 with a TensorFlow backend. Furthermore, we resized images to

242 299×299 pixels in width and height with the "nearest" interpolation method from the

243 python Pillow library. We kept the images in RGB as for some specimens color could be

244 important, giving them 3 pixels in depth. In the end, input was formed as

245 $n \times 299 \times 299 \times 3$ with n as batch number.

246

RESULTS

247 We configured the model to train for a maximum of 200 epochs if not stopped early.
248 The batch size was 32 and the iterations per epoch were defined as the number of images
249 divided by batch size, making sure the model processes all training data each epoch. We
250 programmed the model to stop training if the model did not improve for 50 continuing
251 epochs (due to early stopping) to prevent overfitting. Model improvement is defined as a
252 decline in the loss function for the validation set. We programmed learning rate to decrease
253 with a factor of approximately 0.1 if the model did not improve for 25 continuing epochs.
254 During training, weights were saved for the best model and at the final epoch. Lastly,
255 training, validation and test accuracy and top 3 accuracy were saved after training. Top- n
256 accuracies, (commonly used with $n = 1, 3, 5, 10$), are accuracies that show if any of the n
257 highest probability answers match the true label. The above settings were applied to all
258 experiments.

259

Shot type training

260 In all shot type experiments, validation top 1 and top 3 accuracy rapidly increased
261 the first few epochs and after around 50 – 75 epochs the models converged to an accuracy
262 plateau (Figure 6 on page 32). During training, the learning rate was reduced by factor 10
263 at epoch 47 for dorsal, epoch 66 and 99 for head, epoch 54 and 102 for profile, and epoch
264 50 and 80 for multi-view. At these accuracy plateaus, the models practically stopped
265 improving, so early stopping ceased training at epoch 100, 122, 125, and 104 epochs for
266 dorsal, head, profile, and stitched, respectively. Training usually completed in three and a
267 half hours to four and a half hours, depending on the experiment.

268

Unclean data test results Test accuracy on *top97species_Qmed_def* reached 65.17%,

269 78.82%, and 66.17% for dorsal, head, and profile views, respectively (Table 2 on page 35).

270 Top 3 accuracy reached 82.88%, 91.27%, and 86.31% for dorsal, head, and profile view,

respectively. Genus accuracy reached 82.58%, 93.98%, and 86.94% for dorsal, head, and profile view, respectively. Top 1, top 3 and genus accuracies were obtained directly after training where the model was in its validation accuracy plateau. Therefore, these accuracies do not represent the best model, of which the accuracies are shown later.

Clean data test results Test accuracy on *top97species_Qmed_def_clean* reached 63.61%, 78.55%, and 68.75% for dorsal, head and, profile views, respectively (Table 2 on page 35). Top 3 accuracy reached 81.65%, 91.24%, and 86.31% for dorsal, head, and profile view, respectively. Genus accuracy reached 82.87%, 92.45%, and 87.20% for dorsal, head, and profile view, respectively. Top 1, top 3 and genus accuracies were obtained directly after training where the model was in its validation accuracy plateau. Therefore, these accuracies do not represent the best model, of which the accuracies are shown in the section below.

During training on *top97species_Qmed_def_clean*, the model with the lowest validation loss function was saved at the lowest loss. This model was viewed as the best model, as the error between training and validation was at its lowest, instead of picking the model based on the validation accuracy. The lowest loss model will represent a more robust model than the previous models with higher validation loss, despite having slightly higher validation accuracy. Using the lowest loss model on the test data of *top97species_Qmed_def_clean*, accuracy reached 61.77%, 78.25%, and 67.26% for dorsal, head, and profile view, respectively (Table 2 on page 35). Top 3 accuracy reached 80.12%, 89.73%, and 86.31% for dorsal, head, and profile view, respectively. Genus accuracy reached 79.52%, 93.66%, and 86.90% for dorsal, head, and profile view, respectively.

Breaking down the top 1 prediction for the lowest loss models shows that most of the predictions were correct. To visualize the classification successes and errors we constructed confusion matrices using the true and predicted labels (Figure 7 on page 33). A bright yellow diagonal line indicates that most of the species were classified correctly.

297 *Multi-view test results* An accuracy of 64.31% was reached on the
298 *top97species_Qmed_def_clean_multi* test set (Table 2 on page 35). Top 3 accuracy reached
299 83.69% and genus accuracy 85.85%. Stitched validation accuracy increased the most
300 uniform of all shot type approaches, before reaching a plateau after roughly 50 epochs
301 (Figure 6 on page 32).

302 *Worker only data results*

303 Accuracy for *top97species_Qmed_def_clean_wtest* reached 64.39%, 81.00%, and
304 69.42% for dorsal, head, and profile views, respectively (Table 2 on page 35). Top 3
305 accuracy reached 82.58%, 92.47%, and 87.50% for dorsal, head, and profile view,
306 respectively. Genus accuracy reached 84.47%, 96.42%, and 90.68% for dorsal, head, and
307 profile view, respectively. Head genus accuracy was the highest accuracy found in all
308 experiments.

309 We see that the accuracies go up, but the test set also becomes smaller. To compare
310 this, we took worker accuracy and calculated reproductive accuracy. The head shot type
311 reproductives reached an accuracy of 65.40%, while for workers accuracy reached 81.00%, a
312 difference of 15.60% (Table 3 on page 36). This difference is much larger than for the other
313 shot types; dorsal: 4.04% and profile: 4.88%.

314 *RMNH collection test results*

315 Accuracy for *statia2015_rmnh* reached 17.86%, 14.29%, and 7.14% for dorsal, head,
316 and profile views, respectively (Table 2 on page 35). Top 3 accuracy reached 60.71%,
317 21.43%, and 25.00% for dorsal, head, and profile view, respectively. Genus accuracy
318 reached 21.43%, 25.00%, and 14.29% for dorsal, head, and profile view, respectively. This is
319 the only case where genus accuracy is substantially lower than the top 3 accuracy. Profile
320 top 1 accuracy was the lowest accuracy found in all experiments.

321

DISCUSSION

322 We present an image-based ant classification method with 61.77% – 81.00%
323 accuracy for different shot types. We processed the input for training in different ways and
324 with test data including a worker-only and an AntWeb protocol-deviant test set.
325 Consistently throughout our experiments, shot type accuracies were found to rank from
326 low to high accuracy in the same order: dorsal → profile → head. The head shot type
327 predominantly outperformed dorsal, profile, and stitched in accuracy by about ten
328 percentage points most of the time, perhaps due to the fact that this shot type is more
329 protocol stable. An additional explanation may be that discriminating characters are more
330 concentrated in the head in some ant groups. The combined, stitched image view did not
331 greatly increase accuracy, as the head shot type outperformed the stitched view by 6.04% –
332 7.58%. A not so much curious result, as the combination of multiple views in one image is
333 the most naive way of approaching a multi-view learning problem (Zhao et al. 2017). Other
334 approaches on a multi-view problem (discussed in Section: Recommendations for future
335 work) would most probably have higher accuracies. Genus accuracy reached 79.52% –
336 96.42%, which is approximately as accurate as the top 3 accuracy (80.12% – 92.47%),
337 sometimes slightly above it. It is, however, important to note that the CNN has no
338 understanding of what a genus is, because it selects the label *genus-species* from among a
339 flat list. Top 3 accuracy is preferred over genus accuracy as this will show only three
340 options, of which one is correct, where a genus accuracy could still have over 20 potential
341 species.

342 Looking at the confusion matrices (Figure 7 on page 33) outliers can best be
343 explained as specimens that are morphological-wise very comparable. This is especially the
344 case in *Camponotus*, *Crematogaster* or *Pheidole*, which have a lot of species in the dataset
345 (14, 8, and 17, respectively). In contrast, just eight other genera have two to six genera in
346 the dataset and the rest only one. And because the species in the confusion matrices are
347 alphabetically sorted on genus, false predictions near the yellow diagonal line are most of

348 the time found within the correct genus for these three big genera. Therefore we speculate
349 that inter-genus features are better distinguished than intra-genus features.

350 Because the majority of specimens are workers, there is most probably a bias in
351 learning the workers from a species. We therefore speculate that the model has acquired an
352 improved understanding and representation of workers. However, accuracy for workers did
353 increase only slightly, when reproductives were removed from the test set. We see a slight
354 increase in dorsal and profile worker accuracy over reproductives accuracy, but the increase
355 is small. The only noticeable and interesting increase is for the head shot type, where
356 workers were classified 15.60% more accurate (Table 3 on page 36). We still see a slight
357 increase in dorsal and profile worker accuracy over reproductives accuracy, but the increase
358 is small. It seems that discriminating workers from reproductives is best performed using
359 the head. This could have something to do with ocelli, only present on heads of
360 reproductives, causing trouble.

361 The image number threshold for the species in this data set was 68 images, which is
362 approximately 23 images per shot type. That accounts for 16 images in the training set,
363 which nonetheless achieved good accuracy. This means that the threshold could potentially
364 be lower, and thus more species (with fewer than 68 images) could be incorporated.
365 However, more species (classes) will also complicate training and test accuracy.

366 One of the biggest improvements in accuracy can be made by increasing the data
367 and thus reducing variance (training error) and overfitting. The current data shows a much
368 skewed, long tailed, distribution with the first two species containing over 10% of the total
369 number of images. Furthermore, only *C. maculatus* and *Pheidole megacephala* (Fabricius,
370 1793) had over 100 stitched images out of 3,322 in total. Also important when expanding
371 the image set is adding male and queen specimens so the classifier has improved learning of
372 these castes. Despite the fact that Bolton (2003) provided the first big overview for male
373 ant taxonomy, at this moment 22% of extant species still have their male castes unknown,
374 because males are usually only found incidentally. As males have been found to be

375 important factors in a colony and not just sperm vessels (Shik et al. 2012), it is important
376 to include these underrepresented specimens in automatic ant identification.

377 Results are not shown, but species in a species complex (i.e. species with
378 subspecies) did not complicate training and did not cause accuracy problems. This was
379 measured using the F_1 -score, calculated as the harmonic mean of precision and recall.
380 With an increasing number of species in a complex, the F_1 -score did not increase or
381 decrease significantly; variation in data could not be explained by the linear relation.

382 Of interesting note is the labeling of this data set, as this was not managed by the
383 author. Identifications and labels were directly taken from AntWeb, assuming that they
384 were correct. However, there is always a chance that identifications are less accurate and
385 certain as expected (e.g. Boer (2016)), despite being a by-expert-labeled data set. Reality
386 is that ant identification is more complex work than labeling a cat and dog dataset for
387 example.

388 Despite some obstacles and points for improvement, we have shown that processing
389 data in different ways influences test results in different ways. In this article we
390 demonstrated that it is possible to classify ant species from images with decent accuracy.

391 *Recommendations for future work*

392 To the best of our knowledge, this is the first time ants were classified to species
393 level using computer vision, which also means that there is a lot to improve. In this section
394 we will discuss some possible improvements for future research in the form of
395 recommendations.

396 To start, focus should lie on creating benchmark data set that is easy to enlarge
397 and improve. To do that, first it is important to find the image threshold for the model to
398 learn a species, which could differ per genera and species. Finding this number would shift
399 the focus to photographing species below the threshold in reaching the threshold. To also
400 increase the data set in the near future, specimens from natural history museums ant

401 collections should be photographed, as it would be less time and cost expensive than
402 collecting new specimens. These existing specimens are most likely already following
403 AntWeb mounting standards. Hopefully this could also solve the skewed image distribution
404 and add more three shot type specimens. In the end, this data set could serve as
405 benchmark data for automated insect identification, and then research focus can shift to
406 accuracy-improving efforts.

407 One of these efforts could be the incorporation of a hierarchical system, where the
408 model classifies on different taxonomic levels as Pereira et al. (2016) did with orchids
409 (Asparagales: Orchidaceae). In an effort to do this, one could do this in a series of multiple
410 CNNs (e.g. first subfamily, then genus, then species), but also in three parallel CNNs,
411 learning simultaneously. However, for this we first need to work on a (phylogenetic) tree
412 and molecular data, which is a different study itself. Moreover, there is also the option to
413 classify on caste, before classifying species, using a caste-trained CNN, and then make use
414 of specialized workers, males and queen trained CNNs.

415 An other option is to incorporate metadata; e.g. biogeographic region, caste,
416 country, collection locality coordinates, or even body size (using the included scale bar on
417 images). Metadata could be very important, especially for species that are endemic to a
418 specific region. Metadata could provide underlying knowledge of the characteristics. Most
419 of this information is already present on AntWeb and ready for use.

420 In order to improve the multi-view approach, multiple solutions have been tried
421 (Zhao et al. 2017). A first option is to try is using just one CNN with all images as input
422 and with the addition of catalog number as a label will. The next option could be to train
423 three shot type CNNs parallel and combine the output. The output can be processed as
424 the average of three shot type predictions, or by using the highest prediction. It is also
425 possible to overlay the three images and take the average pixel values in order to create an
426 average single input image of a specimen.

427 Furthermore, as results have shown, it is very important to have the same mounting

428 procedure and photographing protocol to get a uniform set of images. Difference in dried
429 and alcohol material is most likely very important, but other details like type of glue,
430 background, and zoom could potentially be important and will have to be standardized.
431 Also to get high-detail images, the use of good image stacking software and high-resolution
432 cameras is very important. Therefore, the recommendation is to follow the, already widely
433 used, AntWeb protocol (AntWeb.org 2018).

434 In the end, research like this could assist taxonomists, natural history museums,
435 and researchers to achieve higher taxonomic completeness, better collections and therefore
436 improve research. But for general use the code should further be developed in an easy to
437 use application. A functioning application with high accuracy could reduce costs, time, and
438 energy during everyday identification work (Gaston et al. 2004). However, bear in mind
439 that an application like this is not aimed for use in the field and there is still skill required
440 in collecting, mounting and photographing specimens. Nonetheless, we would like to argue
441 that automated species identification from images using a CNN has high potential.
442 Research in this subject should be continued, and even though DL still has some obstacles
443 to overcome (Marcus 2018), it has already advanced a lot (Guo et al. 2016; Wäldchen et al.
444 2018).

445 ACKNOWLEDGEMENTS

446 We thank AntWeb and its community for providing the necessary data, images and
447 information to make this research possible. We would also like to thank Laurens Hogeweg
448 for all the help and discussions on neural networks and the programming that goes with it.
449 Furthermore, we would like to thank dr. Martin Rücklin for use of computers and guidance
450 in the computer lab.

REFERENCES

19

451 SUPPLEMENTARY MATERIAL

452 Programming code and documentation is available for open access (MIT licensed)
453 and published on URL: github.com/naturalis/FormicID.

454 Data available from the figshare repository:

455 <https://doi.org/10.6084/m9.figshare.6791636.v4>

456 REFERENCES

457 Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A,
458 Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y,
459 Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S,
460 Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K,
461 Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P,

462 Wattenberg M, Wicke M, Yu Y, and Zheng X. 2016. TensorFlow: Large-Scale
463 Machine Learning on Heterogeneous Distributed Systems. arXiv: 1603.04467.

464 Affonso C, Rossi ALD, Vieira FHA, and Carvalho ACPdLF de. 2017. Deep learning for
465 biological image classification. Expert Syst Appl. 85: 114–122. DOI:
466 [10.1016/j.eswa.2017.05.039](https://doi.org/10.1016/j.eswa.2017.05.039).

467 Andel T van, Hoorn B van der, Stech M, Arostegui SB, and Miller J. 2016. A quantitative
468 assessment of the vegetation types on the island of St. Eustatius, Dutch Caribbean.
469 Glob Ecol Conserv. 7: 59–69. DOI: [10.1016/j.gecco.2016.05.003](https://doi.org/10.1016/j.gecco.2016.05.003).

470 Andersen AN. 1997. Using ants as bioindicators: Multiple issues in ant community ecology.
471 Conserv Ecol. 1(1): 1–17.

472 Andersen AN, Hoffmann BD, Muller WJ, and Griffiths AD. 2002. Using ants as
473 bioindicators in land management: simplifying assessment of ant community
474 responses. J Appl Ecol. 39(1): 8–17. DOI: [10.1046/j.1365-2664.2002.00704.x](https://doi.org/10.1046/j.1365-2664.2002.00704.x).

475 AntWeb.org. 2017(a). AntWeb. URL: <http://www.antweb.org> (visited on 01/22/2017).

- 476 AntWeb.org. 2017(b). AntWeb API (version 2). URL: <http://www.antweb.org/api/v2/>
477 (visited on 01/22/2017).
- 478 —— 2017(c). AntWeb API (version 3). URL: <https://www.antweb.org/api.do> (visited
479 on 01/22/2017).
- 480 —— 2018. AntWeb Participation. URL: <https://www.antweb.org/documentation.do>
481 (visited on 07/16/2018).
- 482 Barré P, Stöver BC, Müller KF, and Steinhage V. 2017. LeafNet: A computer vision
483 system for automatic plant species identification. *Ecol Inform*. 40(December 2016):
484 50–56. DOI: [10.1016/j.ecoinf.2017.05.005](https://doi.org/10.1016/j.ecoinf.2017.05.005).
- 485 Bengio Y, Courville A, and Vincent P. 2012. Representation Learning: A Review and New
486 Perspectives. *35(8)*. arXiv: [1206.5538](https://arxiv.org/abs/1206.5538).
- 487 Boer P. 2016. Are their native Holarctic Lasius and Serviformica ant species in the USA,
488 other than exotic ones? With a key of the North American Lasius s.str. and
489 Chthonolasius subgenera. English. URL:
490 http://www.nlmieren.nl/IMAGES/Nearctic%20Lasius%7B%5C_%7Dspecies.pdf.
- 491 Bolton B. 1994. *Identification Guide to the Ant Genera of the World*. Cambridge, Mass.:
492 Harvard University Press: 232. DOI: [10.1111/j.1365-3113.1995.tb00102.x](https://doi.org/10.1111/j.1365-3113.1995.tb00102.x).
- 493 —— 2003. “Synopsis and classification of Formicidae.” *Synopsis Classif Formicidae*.
494 American Entomological Institute: 370.
- 495 Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, and Joly A. 2017. Going deeper
496 in the automated identification of Herbarium specimens. *BMC Evol Biol*. 17(181):
497 1–14. DOI: [10.1186/s12862-017-1014-z](https://doi.org/10.1186/s12862-017-1014-z).
- 498 Chatfield K, Simonyan K, Vedaldi A, and Zisserman A. 2014. Return of the Devil in the
499 Details: Delving Deep into Convolutional Nets: 1–11. arXiv: [1405.3531](https://arxiv.org/abs/1405.3531).
- 500 Chollet F. 2015. Keras. URL: <https://github.com/fchollet/keras>.

REFERENCES

21

- 501 D. E. Guyer, G. E. Miles, M. M. Schreiber, O. R. Mitchell, and V. C. Vanderbilt. 1986.
502 Machine Vision and Image Processing for Plant Identification. *Trans ASAE*. 29(6):
503 1500–1507. DOI: [10.13031/2013.30344](https://doi.org/10.13031/2013.30344).
- 504 Deng J, Dong W, Socher R, Li LJ, Li K, and Fei-Fei L. 2009. ImageNet: A large-scale
505 hierarchical image database. *2009 IEEE Conf Comput Vis Pattern Recognit*:
506 248–255. DOI: [10.1109/CVPRW.2009.5206848](https://doi.org/10.1109/CVPRW.2009.5206848).
- 507 Dozat T. 2016. Incorporating Nesterov Momentum into Adam. ICLR Work.
- 508 Dyrmann M, Karstoft H, and Midtiby HS. 2016. Plant species classification using deep
509 convolutional neural network. *Biosyst Eng*. 151(2005): 72–80. DOI:
510 [10.1016/j.biosystemseng.2016.08.024](https://doi.org/10.1016/j.biosystemseng.2016.08.024).
- 511 Edwards M and Morse DR. 1995. The potential for computer-aided identification in
512 biodiversity research. *Trends Ecol Evol*. 10(4): 153–158. DOI:
513 [10.1016/S0169-5347\(00\)89026-6](https://doi.org/10.1016/S0169-5347(00)89026-6).
- 514 Ellis D. 1985. Taxonomic sufficiency in pollution assessment. *Mar Pollut Bull*. 16(12): 459.
515 DOI: [10.1016/0025-326X\(85\)90362-5](https://doi.org/10.1016/0025-326X(85)90362-5).
- 516 Fisher BL and Bolton B. 2016. *Ants of Africa and Madagascar: A Guide to the Genera*.
517 Ants of the world series. University of California Press: 503. URL:
518 <https://books.google.nl/books?id=JtUkDQAAQBAJ>.
- 519 Fisher BL and Cover SP. 2007. *Ants of North America: A Guide to the Genera*. University
520 of California Press: 216.
- 521 Francoy TM, Wittmann D, Drauschke M, Müller S, Steinhage V, Bezerra-Laure MAF, De
522 Jong D, and Gonçalves LS. 2008. Identification of Africanized honey bees through
523 wing morphometrics: two fast and efficient procedures. *Apidologie*. 39(5): 488–494.
524 DOI: [10.1051/apido:2008028](https://doi.org/10.1051/apido:2008028).
- 525 Gaston KJ and O'Neill MA. 2004. Automated species identification: Why not? *Philos
526 Trans R Soc B Biol Sci*. 359(1444): 655–667. DOI: [10.1098/rstb.2003.1442](https://doi.org/10.1098/rstb.2003.1442).

- 527 Gerard M, Michez D, Fournier D, Maebe K, Smagghe G, Biesmeijer JC, and De
528 Meulemeester T. 2015. Discrimination of haploid and diploid males of *Bombus*
529 *terrestris* (Hymenoptera; Apidae) based on wing shape. *Apidologie*. 46(5): 644–653.
530 DOI: [10.1007/s13592-015-0352-3](https://doi.org/10.1007/s13592-015-0352-3).
- 531 Glorot X, Bordes A, and Bengio Y. 2011. Deep sparse rectifier neural networks. *AISTATS*
532 '11 Proc 14th Int Conf Artif Intell Stat. 15: 315–323. arXiv: [1502.03167](https://arxiv.org/abs/1502.03167).
- 533 Groc S, Delabie JH, Longino JT, Orivel J, Majer JD, Vasconcelos HL, and Dejean A.
534 2010. A new method based on taxonomic sufficiency to simplify studies on
535 Neotropical ant assemblages. *Biol Conserv*. 143(11): 2832–2839. DOI:
536 [10.1016/j.biocon.2010.07.034](https://doi.org/10.1016/j.biocon.2010.07.034).
- 537 Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, and Lew MS. 2016. Deep learning for visual
538 understanding: A review. *Neurocomputing*. 187: 27–48. DOI:
539 [10.1016/j.neucom.2015.09.116](https://doi.org/10.1016/j.neucom.2015.09.116). URL:
540 <http://linkinghub.elsevier.com/retrieve/pii/S0925231215017634>.
- 541 Held D, Thrun S, and Savarese S. 2015. Deep Learning for Single-View Instance
542 Recognition. arXiv: [1507.08286](https://arxiv.org/abs/1507.08286).
- 543 Hölldobler B and Wilson E. 1990. *The Ants*. English. Cambridge, Mass.: Belknap Press of
544 Harvard University Press.
- 545 Kang SH, Song SH, and Lee SH. 2012. Identification of butterfly species with a single
546 neural network system. *J Asia Pac Entomol*. 15(3): 431–435. DOI:
547 [10.1016/j.aspen.2012.03.006](https://doi.org/10.1016/j.aspen.2012.03.006).
- 548 Kingma DP and Ba J. 2014. Adam: A Method for Stochastic Optimization. arXiv:
549 [1412.6980](https://arxiv.org/abs/1412.6980).
- 550 Krizhevsky A, Sutskever I, and Hinton GE. 2012. ImageNet classification with deep
551 convolutional neural networks. *Adv Neural Inf Process Syst*. 1097–1105. DOI:
552 [10.1145/3065386](https://doi.org/10.1145/3065386).

REFERENCES

23

- 553 LeCun Y, Bengio Y, and Hinton G. 2015. Deep learning. *Nature*. 521(7553): 436–444. DOI:
554 [10.1038/nature14539](https://doi.org/10.1038/nature14539).
- 555 Lee SH, Chan CS, Wilkin P, and Remagnino P. 2015. Deep-Plant: Plant Identification
556 with convolutional neural networks. arXiv: 1506.08425v1.
- 557 Lee SH, Chang YL, Chan CS, and Remagnino P. 2016. Plant identification system based
558 on a convolutional neural network for the lifestlef 2016 plant classification task.
559 CLEF.
- 560 Lin M, Chen Q, and Yan S. 2013. Network In Network. arXiv: 1312.4400.
- 561 Marcus G. 2018. Deep Learning: A Critical Appraisal: 1–27. arXiv: 1801.00631. URL:
562 <http://arxiv.org/abs/1801.00631>.
- 563 Mishkin D, Sergievskiy N, and Matas J. 2017. Systematic evaluation of CNN advances on
564 the Imagenet. arXiv: 1606.02228.
- 565 Mohanty SP, Hughes DP, and Salathé M. 2016. Using Deep Learning for Image-Based
566 Plant Disease Detection. *Front Plant Sci*. 7(September): 1419. arXiv: 1604.03169.
- 567 Mora C, Tittensor DP, Adl S, Simpson AGB, and Worm B. 2011. How many species are
568 there on earth and in the ocean? *PLoS Biol*. 9(8): 1–8. DOI:
569 [10.1371/journal.pbio.1001127](https://doi.org/10.1371/journal.pbio.1001127).
- 570 Moritz C, Richardson KS, Ferrier S, Monteith GB, Stanisic J, Williams SE, and Whiffen T.
571 2001. Biogeographical concordance and efficiency of taxon indicators for
572 establishing conservation priority in a tropical rainforest biota. *Proceedings Biol
573 Sci*. 268(1479): 1875–1881. DOI: [10.1098/rspb.2001.1713](https://doi.org/10.1098/rspb.2001.1713).
- 574 Oliver I and Beattie AJ. 1996. Invertebrate Morphospecies as Surrogates for Species: A
575 Case Study. *Conserv Biol*. 10(1): 99–109. DOI:
576 [10.1046/j.1523-1739.1996.10010099.x](https://doi.org/10.1046/j.1523-1739.1996.10010099.x).
- 577 Pereira S, Gravendeel B, Wijntjes P, and Vos R. 2016. OrchID: a Generalized Framework
578 for Taxonomic Classification of Images Using Evolved Artificial Neural Networks.
579 bioRxiv. DOI: [10.1101/070904](https://doi.org/10.1101/070904).

- 580 Pik AJ, Oliver I, and Beattie AJ. 1999. Taxonomic sufficiency in ecological studies of
581 terrestrial invertebrates. *Aust J Ecol.* 24(5): 555–562. DOI:
582 [10.1046/j.1442-9993.1999.01003.x](https://doi.org/10.1046/j.1442-9993.1999.01003.x).
- 583 Qin H, Li X, Liang J, Peng Y, and Zhang C. 2016. DeepFish: Accurate underwater live
584 fish recognition with a deep architecture. *Neurocomputing.* 187: 49–58. DOI:
585 [10.1016/j.neucom.2015.10.122](https://doi.org/10.1016/j.neucom.2015.10.122).
- 586 Ruder S. 2016. An overview of gradient descent optimization algorithms. *arXiv:*
587 [1609.04747](https://arxiv.org/abs/1609.04747).
- 588 Schuettpelz E, Frandsen P, Dikow R, Brown A, Orli S, Peters M, Metallo A, Funk V, and
589 Dorr L. 2017. Applications of deep convolutional neural networks to digitized
590 natural history collections. *Biodivers Data J.* 5(e21139): e21139. DOI:
591 [10.3897/BDJ.5.e21139](https://doi.org/10.3897/BDJ.5.e21139).
- 592 Shik JZ, Flatt D, Kay A, and Kaspari M. 2012. A life history continuum in the males of a
593 Neotropical ant assemblage: refuting the sperm vessel hypothesis.
594 *Naturwissenschaften.* 99(3): 191–197. DOI: [10.1007/s00114-012-0884-6](https://doi.org/10.1007/s00114-012-0884-6).
- 595 Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R. 2014. Dropout:
596 A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res.* 15:
597 1929–1958.
- 598 Sun Y, Liu Y, Wang G, and Zhang H. 2017. Deep Learning for Plant Identification in
599 Natural Environment. *Comput Intell Neurosci.* 2017: 1–6. DOI:
600 [10.1155/2017/7361042](https://doi.org/10.1155/2017/7361042).
- 601 Szegedy C, Ioffe S, Vanhoucke V, and Alemi A. 2016. Inception-v4, Inception-ResNet and
602 the Impact of Residual Connections on Learning. *arXiv:* 1602.07261.
- 603 Vetter S de. 2016. Image analysis for taxonomic identification of Javanese butterflies.
604 Bachelor thesis. University of Applied Sciences Leiden.

REFERENCES

25

- 605 Wäldchen J, Rzanny M, Seeland M, and Mäder P. 2018. Automated plant species
606 identification - Trends and future directions. *PLOS Comput Biol.* 14(4): e1005993.
607 DOI: [10.1371/journal.pcbi.1005993](https://doi.org/10.1371/journal.pcbi.1005993).
- 608 Wang J, Ji L, Liang A, and Yuan D. 2012a. The identification of butterfly families using
609 content-based image retrieval. *Biosyst Eng.* 111(1): 24–32. DOI:
610 [10.1016/j.biosystemseng.2011.10.003](https://doi.org/10.1016/j.biosystemseng.2011.10.003).
- 611 Wang J, Lin C, Ji L, and Liang A. 2012b. A new automatic identification system of insect
612 images at the order level. *Knowledge-Based Syst.* 33: 102–110. DOI:
613 [10.1016/j.knosys.2012.03.014](https://doi.org/10.1016/j.knosys.2012.03.014).
- 614 Ward PS. 2007. Phylogeny, classification, and species-level taxonomy of ants
615 (Hymenoptera: Formicidae). *Zootaxa.* 563(1668): 549–563.
- 616 Watson AT, O'Neill MA, and Kitching IJ. 2004. Automated identification of live moths
617 (Macrolepidoptera) using digital automated identification System (DAISY). *Syst*
618 *Biodivers.* 1(3): 287–300. DOI: [10.1017/S1477200003001208](https://doi.org/10.1017/S1477200003001208).
- 619 Weeks PJ, O'Neill MA, Gaston KJ, and Gauld ID. 1999. Automating insect identification:
620 Exploring the limitations of a prototype system. *J Appl Entomol.* 123(1): 1–8. DOI:
621 [10.1046/j.1439-0418.1999.00307.x](https://doi.org/10.1046/j.1439-0418.1999.00307.x).
- 622 Weeks PJD and Gaston KJ. 1997. Image analysis, neural networks, and the taxonomic
623 impediment to biodiversity studies. *Biodivers Conserv.* 6(2): 263–274. DOI:
624 [10.1023/a:1018348204573](https://doi.org/10.1023/a:1018348204573).
- 625 Weeks P, Gauld JD, David II, Gaston KJK, and O'Neill MAM. 1997. Automating the
626 identification of insects: a new solution to an old problem. *Bull Entomol Res.*
627 87(02): 203–211. DOI: [10.1017/S000748530002736X](https://doi.org/10.1017/S000748530002736X).
- 628 Wen C and Guyer D. 2012. Image-based orchard insect automated identification and
629 classification method. *Comput Electron Agric.* 89: 110–115. DOI:
630 [10.1016/j.compag.2012.08.008](https://doi.org/10.1016/j.compag.2012.08.008).

- 631 Wong SC, Gatt A, Stamatescu V, and McDonnell MD. 2016. Understanding data
632 augmentation for classification: when to warp? arXiv: 1609.08764.
- 633 Yang HP, Ma CS, Wen H, Zhan QB, and Wang XL. 2015. A tool for developing an
634 automatic insect identification system based on wing outlines. Sci Rep. 5(1): 12786.
635 DOI: 10.1038/srep12786.
- 636 Yosinski J, Clune J, Bengio Y, and Lipson H. 2014. How transferable are features in deep
637 neural networks? arXiv: 1411.1792.
- 638 Zhao J, Xie X, Xu X, and Sun S. 2017. Multi-view learning overview: Recent progress and
639 new challenges. Inf Fusion. 38: 43–54. DOI: 10.1016/j.inffus.2017.02.007.

FIGURES

27

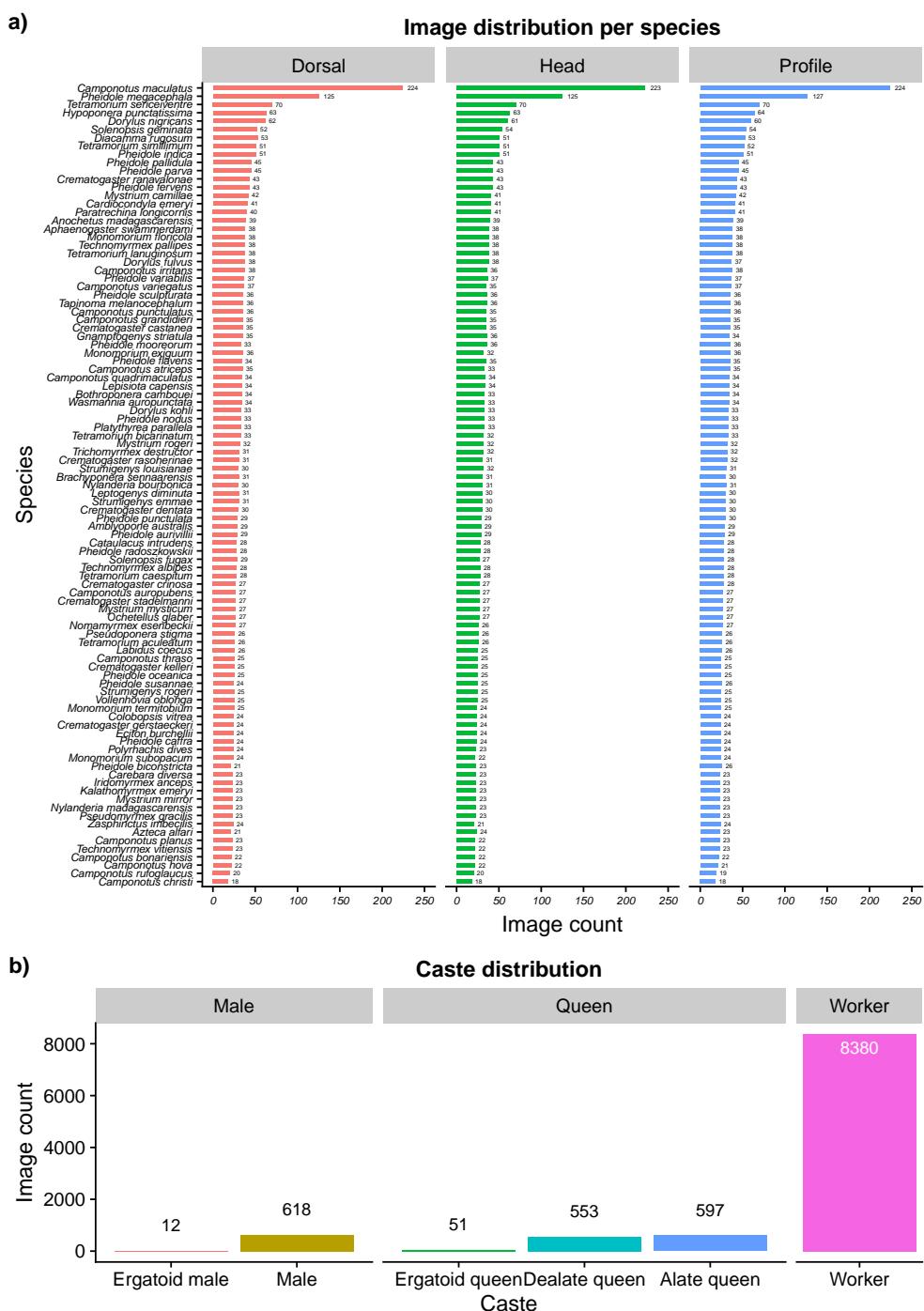


Figure 1. a) Histogram showing the ranked distribution for the 97 most imaged species per shot type (dorsal in red, head in green and profile in blue) for *top97species_Qmed_def*. Species are ranked for the combined shot type image count. Combined image counts ranges from 671 images for *Camponotus maculatus* (Fabricius, 1782) to 54 images for *Camponotus christi* (Forel, 1886). b) Histogram showing the image distribution for the different castes in *top97species_Qmed_def*.

FIGURES



Figure 2. Sample images from *top97species.Qmed.def* showing the diversity in species, shot types, mounting, background, and specimen quality. The images have not been preprocessed. Images were downloaded from AntWeb.

FIGURES

29



Figure 3. Sample image of a stitched image of the dorsal, head and, profile shot type for a *Wasmannia auropunctata* (Roger, 1863) worker (casent0171093). This image has not been preprocessed. Photo by Eli M. Sarnat / URL: <https://www.AntWeb.org/specimenImages.do?name=casent0171093>. Image Copyright AntWeb 2002 - 2018. Licensing: Creative Commons Attribution License.



Figure 4. Example of random data augmentation on a medium quality head view image of a worker of *Eciton burchellii* (Westwood, 1842) (casent0009221). These images have been preprocessed and resized before augmentation. Original photo by / URL: <https://www.AntWeb.org/bigPicture.do?name=casent0009221&shot=h&number=1>. Image Copyright AntWeb 2002 - 2018. Licensing: Creative Commons Attribution License.

Schematic diagram of *Inception-ResNet-V2*

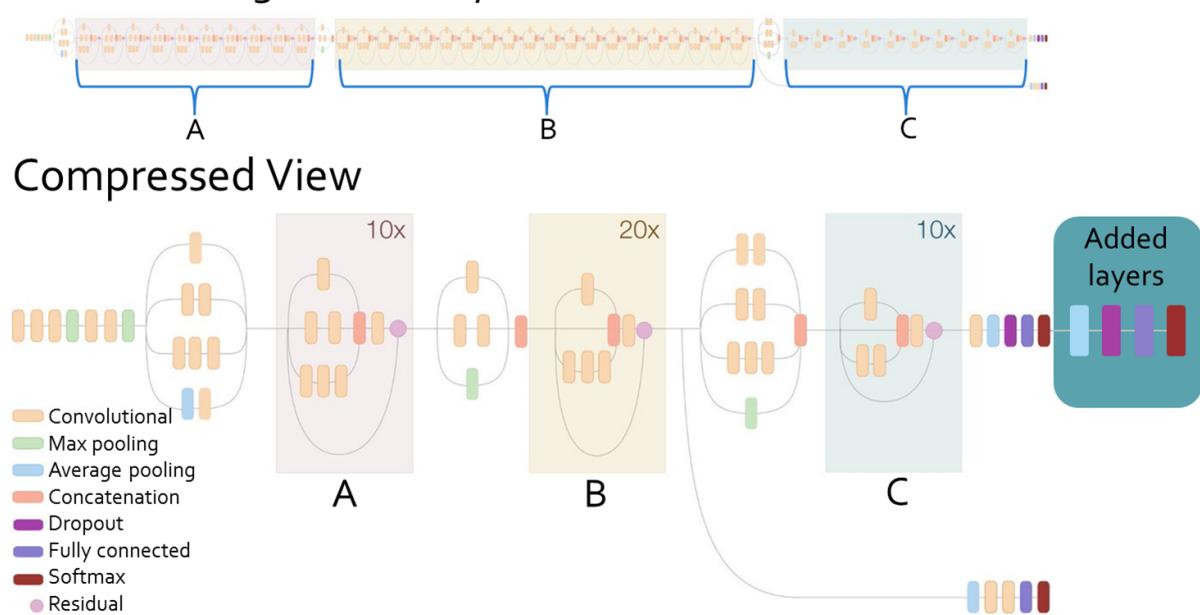


Figure 5. A modified version of Inception-ResNet-V2 (Szegedy et al. 2016) was used as the classifying model. It is built using 3 main building blocks (block A, B and C), each with its own repeating layers. On top of the shown network, four top layers were added, in order: global average pooling layer, dropout, fully connected layer with ReLU, and a fully connected softmax layer. Image is adjusted from:

<https://ai.googleblog.com/2016/08/improving-inception-and-image.html>.

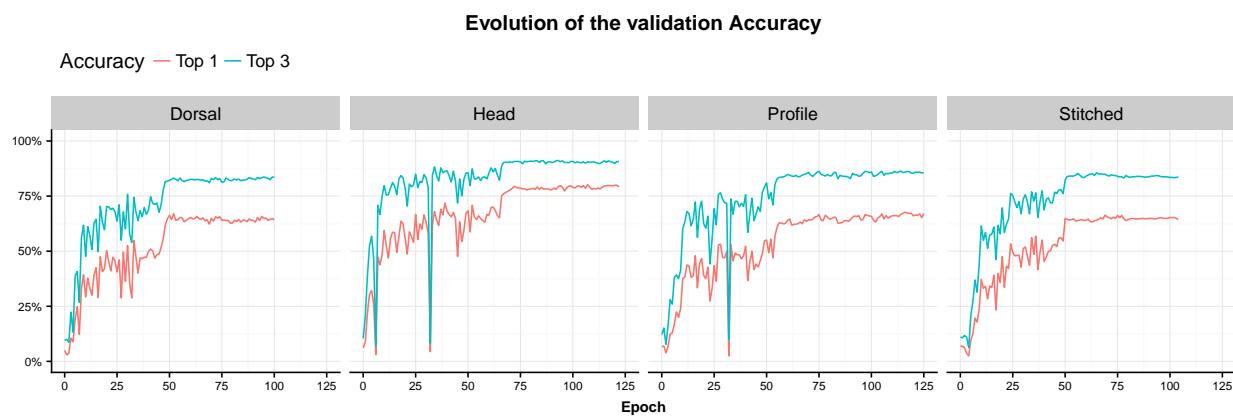


Figure 6. Evolution of the validation accuracy for *top97species_Qmed_def_clean* for different shot types during training (in red: top 1 accuracy, in blue: top 3 accuracy). Where the line ends, training was ceased due to early stopping. From left to right: dorsal, head, profile and stitched shot type.

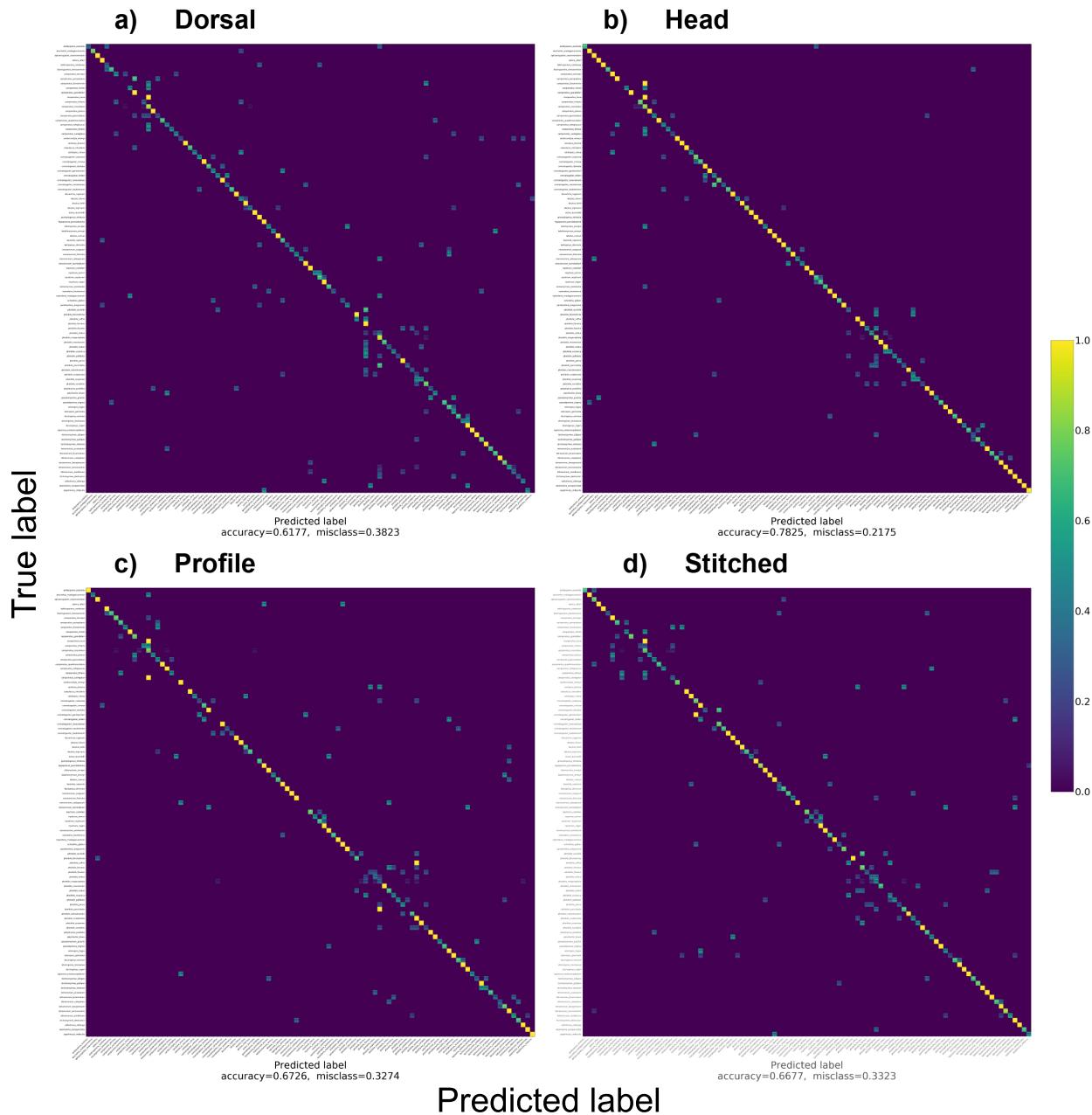


Figure 7. Confusion matrices showing the true label (x-axis) and predicted label (y-axis) for the dorsal (a), head (b), profile (c) and stitched (d) test sets. Each row and column represents a species. Classification accuracies are 0.6177 (a), 0.7825 (b), 0.6726 (c) and 0.6677 (d) (see also Table 2 on page 35). Most confusion was found within large genera like *Camponotus* or *Pheidole*. Confusion matrices were made using the model with the lowest validation loss trained on *top97species_Qmed_def_clean*. Prediction accuracy is indicated by color; from 1.0 – correct (yellow) to 0.0 – incorrect (blue). Numbers in the cells are normalized to [0, 1] to show the prediction accuracy; zeroes are not shown (best viewed on computer).

Table 1. *Image distribution for different data sets for training, validation and test sets for 70%, 20% and 10%, respectively. top97species_Qmed_def_clean_wtest and statia2015_rmnh have no training and validation images, because they are test data sets.*

| | Shot type | def | def_clean | def_clean_multi | def_clean_wtest | statia2015_rmnh |
|------------|-----------|-------|-----------|-----------------|--------------------|-----------------|
| Specimens | | 3,437 | 3,407 | 3,322 | 2,843 ^a | 28 |
| Total | Dorsal | 3,405 | 3,362 | - | 264 | 28 |
| | Head | 3,385 | 3,378 | - | 279 | 28 |
| | Profile | 3,421 | 3,377 | - | 278 | 28 |
| | Stitched | - | - | 3,322 | - | - |
| Training | Dorsal | 2,381 | 2,354 | - | 0 | 0 |
| | Head | 2,364 | 2,358 | - | 0 | 0 |
| | Profile | 2,392 | 2,362 | - | 0 | 0 |
| | Stitched | - | - | 2,322 | - | - |
| Validation | Dorsal | 691 | 681 | - | 0 | 0 |
| | Head | 689 | 689 | - | 0 | 0 |
| | Profile | 692 | 679 | - | 0 | 0 |
| | Stitched | - | - | 675 | - | - |
| Test | Dorsal | 333 | 327 | - | 264 | 28 |
| | Head | 332 | 331 | - | 279 | 28 |
| | Profile | 337 | 336 | - | 278 | 28 |
| | Stitched | - | - | 325 | - | - |

^a 2,843 specimens were marked as valid worker specimens and, therefore, were possible specimens for the worker only test set.

FIGURES

35

Table 2. *Test accuracies for different data sets and all shot types. Top 1, top 3 and genus accuracy results are shown.*

| Accuracy | Shot type | <i>def</i> | <i>def_clean</i> | Best model | <i>def_clean_multi</i> | <i>def_clean_wtest</i> | <i>Statia2015_rmnh</i> |
|----------|-----------|------------|------------------|------------|------------------------|------------------------|------------------------|
| Top 1 | Dorsal | 65.17% | 63.61% | 61.77% | - | 64.39% | 17.86% |
| | Head | 78.82% | 78.55% | 78.25% | - | 81.00% | 14.29% |
| | Profile | 66.17% | 68.75% | 67.25% | - | 69.42% | 7.14% |
| | Stitched | - | - | - | 64.31% | - | - |
| Top 3 | Dorsal | 82.88% | 81.65% | 80.12% | - | 82.58% | 60.71% |
| | Head | 91.27% | 91.24% | 89.73% | - | 92.47% | 21.43% |
| | Profile | 86.31% | 86.31% | 86.31% | - | 87.50% | 25.00% |
| | Stitched | - | - | - | 83.69% | - | - |
| Genus | Dorsal | 82.58% | 82.87% | 79.52% | - | 84.47% | 21.43% |
| | Head | 93.98% | 92.45% | 93.66% | - | 96.42% | 25.00% |
| | Profile | 86.94% | 87.20% | 86.90% | - | 90.68% | 14.29% |
| | Stitched | - | - | - | 85.85% | - | - |

Table 3. *Correct and incorrect predictions, and top 1 test accuracies for workers and reproductives on top97species_Qmed_def_clean. Reproductive count and accuracy is calculated from the differences in correct and incorrect predictions between top97species_Qmed_def_clean and top97species_Qmed_def_clean_wtest. Worker accuracy is taken from Table 2 on page 35.*

| Shot type | Workers | | | Reproductives | | |
|-----------|---------------------|-----------------------|----------|---------------------|-----------------------|----------|
| | Correct predictions | Incorrect predictions | Accuracy | Correct predictions | Incorrect predictions | Accuracy |
| Dorsal | 170 | 94 | 64.39% | 38 | 25 | 60.34% |
| Head | 226 | 53 | 81.00% | 34 | 18 | 65.40% |
| Profile | 193 | 85 | 69.42% | 38 | 20 | 65.54% |