Investigating image-based species identification for birds in the wildlife trade

Sicily Bambini Fiennes

Durrell Institute of Conservation and Ecology

School of Anthropology and Conservation

University of Kent, Canterbury

Thesis submitted for the degree of Msc by Research in Biodiversity Management

September 2021

Word count: 26,273

Investigating image-based species identification for birds in the wildlife trade

Sicily Bambini Fiennes

Supervised by:

Dr. David L. Roberts

Professor Julio Hernandez-Castro

Acknowledgements

I am very grateful to several people whose support made this project possible. First off, I would like to thank my supervisors David Roberts and Julio Hernandez-Castro, whose insights, expertise and patience allowed me to grow as a researcher. My GPU and brain are still hurting after this thesis!

Another hugely important person in this process has been Toto (Liang Zhijian), my friend and research assistant for part of the data collection in China – you are one of the coolest people I know and a valued peer.

In terms of photo submissions, I would like to thank Jessica Lee, James Eaton and Serene Chng for meeting with me, responding to questions over the year and providing many beautiful photos to train this model. I would also like to thank Sunny Nelson at Lincoln Zoo for advising me how to contact zoos and interact with birds in captive settings.

I would like to thank Harry Marshall for his kindness and willingness to organise action on the songbird extinction crisis in Southeast Asia. This work would also not have been possible without the kindness of my good friend Sam Watts who helped design the Docker image which allowed me to run my code faster.

To my twin sister Fran and my mum, Sine – who have patiently read drafts of my thesis, helped me make my work more readable by a layman or non-specialist audience, and supported me through absolutely everything throughout a difficult year.

Lastly, a huge thanks to Tim Blackburn, Chris Shepherd and Phill Cassey, who have supported me on my bird trade research journey for many years now. To Karlina Indraswari and Nuruliawati Yuwono, both exceptional female Indonesian researchers and super mums. This thesis is dedicated to the both of you as a small contribution to scientific knowledge and monitoring bird trade.

Abstract

In Southeast Asia, songbirds, parrots, owls, woodpeckers, and eagles are in high demand. The depletion of wild bird populations in this region led to the declaration of an Asian Songbird Extinction Crisis in 2017. This trade is megadiverse, and there is a lack of technology to identify the many bird species that appear in wildlife markets. This thesis focuses on bird trade in Indonesia and China, where demand for songbirds is particularly intense. To test if machine learning is viable for this problem, Chapter 2 sets a baseline for human identification error with a matching task requiring same/different decisions for pairings of 19 species. Chapter 3 trained, tested, and compared the capabilities of five deep learning computer vision networks and an ensemble of all networks, using a custom-built data collection, processing, and training pipeline. The best model (using the DenseNet-201 network) achieved a high test prediction accuracy (94.4%), using cross-validation for 37 classes on unseen data. We also demonstrate the effects of the visual dominance of cages (25%, 50% and 75% of images occluded) on test accuracy, precision, recall and the F1-score by artificially placing cage bars in the foreground of images for a subset of 26 species. Chapter 4 builds on Chapters 2 and 3 by comparing the human baseline to the top-performing computer model trained to identify the same species. The computer model performed better than the average human accuracy but worse than the best human score. These results suggest that computers can reliably outperform the average, non-expert human in bird identification tasks. It is hoped that this work will help demystify previous roadblocks for using machine learning to identify birds from pictures in wildlife markets. Image-based machine learning approaches hold great promise for identifying birds and other taxa in highly occluded environments.

Table of contents

Acknowledgements .... **iii**

Abstract .... **iv**

Table of contents .... **v**

List of tables .... **viii**

List of tables in the Appendices .... **ix**

List of figures .... **x**

List of figures in the Appendices .... **xi**

Chapter 1 Introduction **1**

1.1 Aims and objectives 8

Chapter 2 Creating a baseline of human-based identification in the wildlife trade **10**

2.1 Introduction 11

2.2 Methods 14

2.2.1 Ethics 14
2.2.2 Data collection 14
2.2.3 Participant recruitment 14
2.2.4 Survey design 15
2.2.5 Assumptions 17
2.2.6 Calculating metrics of performance 18
2.2.7 Statistical analysis 20
2.2.8 Phylogenetic analysis 21

2.3 Results 23

2.3.1 Testing for outliers 25
2.3.2 Visualising the phylogeny 25

2.4 Discussion 27

2.4.1 Match versus mismatch questions 27
2.4.2 Confidence levels 28
2.4.3 Agreement levels 30
2.4.4 The existence of super-recognisers 31
2.4.5 Phylogenetic pattern of accuracy 31
2.4.6 Limitations 32

2.5 Conclusion 34

Chapter 3 Developing a supervised, multiclass image classification model to identify birds present in the wildlife trade......................................35

3.1 Introduction..36

3.1.1 Exploring the challenges of species identification for birds 40

3.2 Methods...42

3.2.1 Ethics 42
3.2.2 Step 1: Data collection..44
3.2.3 Step 2: Initial species identification and Step 3: Using expert verification...45
3.2.4 Data pre-processing...45
3.2.5 Step 4: Using the MegaDetector, an object detection model..........45
3.2.6 Step 5: Reviewing the MegaDetector output and 6. Cropping photos manually..46
3.2.7 Using a small dataset and the risks of overfitting................................47
3.2.8 Step 7: Using offline data augmentation for class balancing and Step 8. Using online augmentation during training 48
3.2.9 The structure of neural networks 49
3.2.10 Hyperparameter fine-tuning 50
3.2.11 10. Training models which are robust to occlusion.............................51
3.2.12 Using transfer learning to train image-based multiclass classification models..53
3.2.13 Model structure..54
3.2.14 Performance metrics 56
3.2.15 Evaluation metrics 56
3.2.16 Understanding the features used by models to perform identifications..57
3.2.17 Building a Shiny application..57
3.2.18 Using the application...58
3.2.19 Further understanding the effects of the occlusion on image classification..58
3.2.20 Hardware requirements and code release..60

3.3 Results 61

3.3.1 MegaDetector performance...61
3.3.2 Hyperparameter fine-tuning 62
3.3.3 Training different architectures...63
3.3.4 Using stricter evaluation metrics to assess model performance.....67
3.3.5 Applying the Grad-CAM algorithm 71
3.3.6 The Shiny application...73
3.3.7 Testing the effects of occlusion on partitioned datasets 76

3.4 Discussion 78

3.4.1 Applying the MegaDetector...78
3.4.2 Varying other hyperparameters..79
3.4.3 Discussion of transfer learning and experimenting with different architectures. 79

3.4.4 Cross validation of models 83
3.4.5 The performance of the Grad-CAM 83
3.4.6 Model generalisability 83
3.4.7 The impact of occlusion in model performance....84
3.5 Conclusions 86

Chapter 4 Discussion 87

3.5.1 Comparing the performance of humans versus computers 87
3.5.2 What makes an adequate baseline for human identification and what can they be used for?....90
3.5.3 Which features do humans and computers rely on for identification?....91
3.5.4 Building a case for human-in-the-loop approaches to monitor the wildlife trade. 92
3.5.5 Future benefits of the cagedbirdID application....93
3.5.6 Understanding the motivations of end-users of applications and deployment contexts....94
3.5.7 Data collection concerns 95
3.5.8 The potential of multilabel classification....97
3.5.9 Towards market methodologies in the future....98
4.1 Conclusions 100

References....102

Appendix I Chapter 2 Appendix 114

Appendix II Chapter 3 Appendix 115

Appendix II.1 Image inpainting.... 118
Appendix II.2 Predictions from the Shiny app interface.... 124

List of tables

Table 2.1: An explanation of the terms true positive, the true negative, the false positive, and the false negatives rates in the context of our experiment.....................19

Table 2.2: Summary of different metrics relating to all questions, match questions only, and mismatch questions only. 23

Table 2.3: Correlation matrix between accuracy, confidence, and agreement across all questions.. 24

Table 3.1: The different machine learning architectures introduced in the Introduction tested on the TOT_SP_37 dataset..54

Table 3.2: The model performance of a DenseNet-121 architecture with different batch sizes, using Adam as an optimiser with a learning rate of 0. 00001.................... 62

Table 3.3: The model performance of a DenseNet-121 architecture with different optimisers, with TensorFlow default learning rates /arguments, using a consistent batch size of 16.. 62

Table 3.4: The model performance of four neural networks, one transformer, and one ensemble model, with a batch size of 16, using the SGD optimiser and early stopping patience of 3 epochs. 64-65

Table 3.5: The performance of the best model, the DenseNet-201 on, different folds, with data augmentation, random erasing, and a batch size of 16, using the SGD optimiser and early stopping patience of 3 epochs. 68

Table 3.6: The performance of the Vision Transformer on different folds, with data augmentation, random erasing, and a batch size of 16, using the Adam optimiser and early stopping patience of 3 epochs. 69

Table 3.7: The performance of the best model, the DenseNet-201, on datasets with differing degrees of occlusion, using data augmentation, random erasing, and a batch size of 16, with the SGD optimiser and early stopping patience of 3 epochs. 77

List of tables in the Appendices

Table A1.1: The complete data for the MA_MIS_19 dataset. The family, English name, Scientific name, IUCN conservation status, and CITES listing for the nineteen birds used in the match-mismatch experiment in Chapter 2 are compared with a computer model in Chapter 4. 114

Table A2.1: The complete list of species for the TOT_SP_37 dataset. The family, English name, Scientific name, IUCN conservation status, and CITES listing is provided........ 114-116

Table A2.2: The classification report containing precision, recall and F1 score for each of the species in the test set of TOT_SP_37. 121-122

Table A2.3: The training data similarity matrix across the five folds, partitioned from the original TOT_SP_37. There are 6,110 photos in each fold. 123

Table A2.4: The validation data similarity matrix across the five folds, partitioned from the original TOT_SP_37. There are 6,110 photos in each fold. 123

Table A2.5: The test data similarity matrix across the five folds, partitioned from the original TOT_SP_37. There are 6,110 photos in each fold. 123

List of figures

Figure 2:1 An example of a match side-by-side pairing, in the Google Forms survey. Each question was the same, and the Likert scale can be observed underneath. 17

Figure 2.2: A boxplot of all participants' mean, standard deviation, and variance for the match versus mismatch questions. 24

Figure 2.3: A boxplot of the mean, standard deviance, and variance of scores across participants. The four circles at the lower tail of the plot are the outlying scores...... 25

Figure 2.4: A radial phylogeny showing the evolutionary relationships between the nineteen species included in our match-mismatch experiment. 26

Figure 3.1: The methodological data collection and model training pipeline for image classification, created for this study. 43

Figure 3.2: The effect of random erasing when applied to 15 images in a batch, with a probability of random erasing being applied of 0.5. 53

Figure 3.3: The effect of the MegaDetector on cropping an image of multiple caged birds of the species Scaly-breasted Munia, *Lonchura punctulata*.................................61

Figure 3.4: A visualisation of the training process for the best modelwith respect to accuracy and loss. The validation loss was the metric that controlled early stopping, with a pateience of 3 epochs..66

Figure 3.5: The normalised confusion matrix for the best model, the DenseNet-201, on different folds, with data augmentation, random erasing, and a batch size of 16, using the SGD optimiser and early stopping patience of 3 epochs. 70

Figure 3.6: The Grad-CAM image shows the heat map, highlighting the region of 'most interest' to the CNN... 72

Figure 3.7: The performance of the Shiny application applied to a sample of caged, cropped images, along with the top-5 accuracy. 74

Figure 3.8: The performance of the Shiny model applied to a sample of wild, cropped images, along with the top-5 accuracy. 75

Figure 4.1: A) The phylogeny and the associated accuracy per species for the human classification. B) The phylogeny and the associated accuracy per species can be viewed below for the machine learning model classification. 89

List of figures in the Appendices

Figure A2.1: The effect of artificial dying on species appearance, as occasionally seen when Red-billed Starling's, *Spodiospar sericeus* and Scaly-breasted Munia's, *Lonchura punctulata* are offered for sale. 117

Figure A2.2: A sample of cage masks generated from original bird images with cages in the foreground in the image editing software, GIMP.......... 117

Figure A2.3: Cage bars, prison, and transparent fence pngs 118

Figure A2.4: The effect of image inpainting using the EdgeConnect GUI on three different species... 120

Figure A2.5: A screenshot of the GUI of the Shiny application, deployed locally.......124

Figure A2.6: The performance of the Shiny model applied to a sample of uncaged, cropped images, along with the top-5 accuracy. 125

Figure A2.7: The performance of the Shiny model applied to a sample of caged, uncropped images, along with the top-5 accuracy. 126

Figure A2.8: The performance of the Shiny model applied to a sample of wild, uncropped images, along with the top-5 accuracy. 127

Chapter 1 Introduction

Birds play a critical role in the function of almost every ecosystem worldwide, with over 11,000 species of bird acknowledged. Birds themselves are inherently of value, as is all biodiversity, yet they are also of vital importance for human survival. Despite their importance, several bird populations are globally declining from human activities (Şekercioğlu et al., 2016). A 'wicked' problem for global biodiversity has long been the commodification of wildlife into markets (Rittel & Webber, 1973; Silvertown, 2015). This problem has been particularly severe for birds.

Birds involved in the wildlife trade provide direct, consumptive services to humans, such as for food, medicine or ornaments, and to the rest of the ecosystem through seed dispersal, nutrient cycling, pest control and pollination (Michel et al., 2020; Şekercioğlu et al., 2016). Beyond these physical services, all bird species provide cultural services via bird-focussed tourism and inspiration for art and religion (Jepson, 2010; Michel et al., 2020). However, when wild harvesting of birds occurs, the potential for the large-scale removal of species could result in a reduction and eventual loss of these invaluable services (Coleman et al., 2019; Nyffeler et al., 2018; Sheherazade et al., 2019). Scheffers et al. (2019) highlighted a particularly notable trend for birds, in that traded species are more evolutionarily distinct from each other than non-traded species. This non-randomness in bird trade implies high selectivity for specific clades based on evolutionary traits, often aesthetic attributes such as their song (Blackburn et al., 2014) and plumage (Cassey et al., 2004).

The selectivity of bird trade (i.e., targeting parrots for ornamental trade (Pires et al., 2021) and falcons for hunting or as talismans (Wyatt, 2014)) is troubling for several reasons. The removal of wild birds with specialised functional roles could cause a reduction in the provision of certain services. Ecosystem service loss, may in turn, have a cascading effect on other ecosystem services. For example, in a global review of threats to parrot species, Olah et al. (2016) found that forest-dependent species are more likely to be threatened. Parrots are predominantly seed and fruit eaters, which allows for the prediction that the services they provide, such as seed dispersal and pollination, may be reduced by harvesting for wildlife trade. This is of even further concern when we consider that the mortality rate for birds, especially in the illegal trade, is estimated to range from 30% to as high as 90% (Wyatt et al., 2021). Thus, even if individuals are repatriated to the wild, this would not replace the original volume of birds taken.

Scheffers et al. (2019) reported that 45% of all bird species are affected by the wildlife trade. Although this is a contested statistic (Challender et al., 2021), other estimates put the number of species traded from 2,600 to 3,337 from all known species (Blackburn et al., 2014; Butchart, 2008). If demand becomes intense and species are overharvested, trade may switch to a congeneric species with similar traits (Eaton et al., 2015; Harris et al., 2015; Scheffers et al., 2019). This phenomenon may explain why many species are in trade, despite the evident selectivity for particular species or genera. It is seen in many demand hotspots for birds, such as Brazil, China and Indonesia (Daut et al., 2015; Harris et al., 2017; L. Li & Jiang, 2014; Souto et al., 2017).

Consequently, hunting and trapping for wildlife trade is the third-highest threat to birds after habitat disturbance and the presence of invasive species (Harris et al., 2015). Between five and ten million birds are removed annually from the wild for export to pet markets (Carrete & Tella, 2008). The world bird trade is generally biased towards Psittaciformes (20% of all trade) (e.g., parrots, macaws, cockatoos, conures and parakeets) (Blackburn et al., 2014; Iñigo-Elias & Ramos, 1991). The Passeriformes (e.g. tits, thrushes, sparrows, finches, bulbuls, flycatchers and sunbirds), especially the songbirds, are extremely popular (Blackburn et al., 2014). Passeriformes constitute 59% of avian diversity, 30% of which have appeared in trade, comprising 70% of all trade (Blackburn et al., 2014). Historically, parrots are the group that humans have most targeted in the highest volume (Iñigo-Elias & Ramos, 1991).

Parrots are captured for their colourful appearance, intelligence and ability to recognise the human voice (Cassey et al., 2004; Olah et al., 2016; Pires et al., 2021). At least 259 of 355 parrot species have appeared in commercialised trade (Ribeiro et al., 2019). Further, wildlife trade is thought to have contributed to nearly 30% of the 355 species of parrots being currently threatened with extinction (Ribeiro et al., 2019).

In Southeast Asia, there has been an exponential rise in domestic bird trade (Eaton et al., 2015; H. Marshall et al., 2019; Nash, 1993), where more than 1,000 different species have been recorded in trade (Harris et al., 2017). The escalation of bird trade in Southeast Asia has been amplified by several anthropogenic factors facilitating access to birds. Markets with a wide range of species cater to a region with growing population size and increasing demand for wildlife as pets (Nijman, 2010). Habitat conversion, degradation and forest fragmentation have improved access to forests (Bush et al., 2014; Sagar et al., 2021).

In recent years, there has been a rise in bird market surveys globally. This corresponds with considerable research attention to the field of wildlife trade more broadly. Many bird markets have been surveyed across Southeast Asia, with either a geographical or with a species-specific focus; though most of the survey effort has been concentrated on the megadiverse markets of Indonesia, notably on the island of Java (Chng & Eaton, 2016; H. Marshall et al., 2019, 2020). A key challenge for wildlife conservation is therefore, how to monitor and regulate bird trade given its diversity. Additionally, only a few skilled individuals can confidently identify an adequate proportion of species present in the markets.

Indonesia is widely thought to be the current epicentre of global bird trade. Traditionally parrots have been very popular in Indonesia, largely as pets, display or as ornaments (Pires et al., 2021; Setiyani & Ahmadi, 2020). Indonesia hosts 89 species of parrot (Pires et al., 2021); though non-endemic, Southeast Asian parrots are also popular throughout the country (Aloysius et al., 2020). The instigation of a ban by the European Union in 2007 on imports of wild birds caused a substantial decline in live parrot exports from sourcing hotspots such as Indonesia (Cardador et al., 2017; Cooney & Jepson, 2005). Pires et al. (2021) noted a significant overlap between domestic and international trade of certain Indonesian parrot species. Hence, despite bans, local demand has maintained pressure on parrot populations. As of 2021, thirty-four percent of Indonesia's parrot species are still in trade (Pires et al., 2021).

Although parrot trade and its impacts have perhaps been better publicised, there has been a considerable rise in demand for passerines (especially songbirds) in Indonesia. Whilst songbirds have been common pets in Indonesia for several years, the advent of singing competitions and the *kicau mania (*chirping craze) phenomenon in the 21st century has created unprecedented pressure on wild songbird populations. Some estimates suggest that up to 84 million birds are in captivity just on the island of Java (H. Marshall et al., 2019). Other groups such as owls, woodpeckers, and eagles are also traded. Consequently, in 2017, the International Union for the Conservation of Nature (IUCN) declared an Asian Songbird Extinction Crisis (Eaton et al., 2015; Harris et al., 2017; C. Shepherd & Cassey, 2017).

In songbird competitions, birds are assessed on their vocal features such as tempo, tone, and quality (Nash, 1993). Songbird contests take place almost daily in Indonesia, from local to national levels. The prize-winning bird in some competitions can win up to USD 100,000. Although contests are most widespread in Indonesia, they are also popular in Vietnam, Thailand, Taiwan, and Singapore. There is alarming demand for wild-caught songbirds, rooted in the belief that their song is of higher quality (Bergin et al., 2018; Burivalova et al., 2017). Further, only males are entered into competitions, so this creates a genetic skew in the wild.

In Indonesia, the trapping and trading of birds is primarily driven by the demand for pets (H. Marshall et al., 2020). Bird trade in Indonesia is widespread and characterised by a broad range of unique practices, which can be location-specific. For example, cockfighting has deep roots in Bali (Jepson, 2010). In addition, songbird ownership and competitions originated in Java and then dispersed across the archipelago due to migration (Indraswari et al., 2020). Consequently, there are now several bird market hubs across Indonesia. The most diverse and arguably famous is the Pramuka market in Jakarta (Chng et al., 2015), though there are sizeable markets in Eastern and Central Java (Chng & Eaton, 2016), Medan, North Sumatra (C. R. Shepherd, 2006), Denpasar in Bali (Chng et al., 2018), and West Kalimantan (Rentschlar et al., 2018).

Although Indonesia may be the epicentre for bird trade in Southeast Asia, there are also megadiverse markets across China (Dai & Zhang, 2017), which have been shown to host unique trends for species. China has a deep-rooted cultural appreciation for birds, with many unique practices occurring in China, such as 'bird walking', where the bird is walked in its cage by its owner for fresh air and they socialise with other bird owners. There are also trends for using small-bodied tit and thrush species, such as the Siberian Rubythroat, *Calliope calliope* and the Bluethroat, *Luscinia svecica* as companions or pets in China. Likewise, the Japanese Grosbeak, *Eophona personata* is trained to chase seeds and perform other tricks (Liang Zhijian, November 2020, pers. comms). These species, meanwhile, are not commonly observed in the Indonesian trade.

Wildlife market surveys are common for monitoring place-specific markets (Barber-Meyer, 2010; Bušina et al., 2020; Pires, 2015), particularly for birds. Given the sheer diversity of species and several unknowns such as market turnover and drivers of market composition, surveys are often only 'snapshots' of trade (Chng & Eaton, 2016; Eaton et al., 2017) or the so-called 'tip of the iceberg' (Bušina et al., 2020). Species are generally recorded by counting openly displayed individuals offered for sale at individual shops or stalls, using the widely practised Direct Counting Method (DCM) (Bušina et al., 2020). Photos can be taken to identify species during surveys, verify species identifications post-survey, or act as evidence later.

Roberts and Hinsley, (2020) outlined several dimensions of wildlife trade that are often investigated (i.e., the taxa, geographic origin, and production origin) in research. Identification of the 'taxonomic unit', i.e., species is challenging for megadiverse groups such as birds, as hundreds of species are involved. Another dimension of trade is the geographic origin of species, which has important implications for national legislation. The location where a bird is recorded, in a market, for example, can aid identification based on existing knowledge of their distribution. Finally, its production origin is vital due to overlaps between illegally wild-caught birds and legal captive-bred individuals.

The 'taxonomic unit' challenge has been highlighted in seizure and market studies performed by Indraswari et al. (2020) and Rentschlar et al. (2018). Indraswari et al. (2020) collated seizure data from news reports across Indonesia. They found that of the 132,945 birds recorded, over one-third (36%) could not be identified to species or even family level. In their study on bird markets in Kalimantan, Borneo, Rentschlar et al. (2018) could not identify 25% of the 25,298 individuals to species level. Rentschlar et al. (2018) attributed this inability to identify the species as reportedly due to the high number of individuals from look-alike species complexes.

Consequently, the taxonomic diversity of birds poses challenges in quantifying species diversity at markets, and training practitioners in species identification. For birds, species identification is principally based on visual recognition and the geographic location of sale, if known. Thus, image-based approaches for accurate species identification could hold great promise for addressing this trade. Given the ubiquity of mobile phones, we can now integrate powerful artificial intelligence applications into small devices such as phones and microcontrollers. Subsequently, automated species identification using photos can now be conducted in the field.

Accurate field identification of illegally traded wildlife and their associated products is a challenge for conservationists and law enforcement agencies alike (Kretser et al., 2015). The use of technology to address data gaps in the wildlife trade has become more common in recent years. For instance, this has been done using mobile applications to combat illegal wildlife trade, such as the Wildlife Witness app, where members of the public can report illegal wildlife trade by taking a photo and the GPS location of an incident. Wildlife Witness is similar in rationale to other apps, such as the WildScan app and the Wildlife Conservation Society mobile app, which use decision-tree tool logic to assess the legality of a specimen and thus detect wildlife crime in China, Vietnam and Afghanistan (Kretser et al., 2015).

Despite the emergence of wildlife crime technology challenges, apps frequently suffer from a lack of continuous funding and poor uptake by law enforcement (Joanny, 2020). Although apps exist, there is no open-source technology or code to accompany them. Further, the majority of these are mobile phone-based. Species identification technology needs to have visually appealing user interfaces which facilitate taxon identification with a minimum number of steps (Farnsworth et al., 2013). Apps such as those introduced by Kretser et al. (2015) may confuse users with an extensive process of elimination to narrow the potential pool of species candidates at each step. Stringham et al. (2021) highlight the lack of applications of image classification tools to web data of the wildlife trade. There is currently a shortage of evidence regarding whether wildlife can be accurately recognised from photos taken in markets and a lack of scientific literature providing such evidence.

In academic research, Di Minin et al. (2019) provided a research framework employing machine learning to investigate illegal wildlife trade on social media platforms. Their framework has three stages: mining, filtering, and accurate identification. Despite the existence of such frameworks, there has been little published on applications of machine learning to species identification in the field of wildlife trade. The few existing exceptions concern the illegal elephant ivory trade. Xu et al. (2019), used text-based machine learning models to detect the illict trade of elephant ivory and pangolin amongst Mandarin-speaking users on Twitter. Hernandez-Castro and Roberts (2015) developed an automated system to detect potentially illegal elephant ivory items for sale on eBay on post metadata.

Machine learning, specifically supervised machine learning, using convoluted neural networks (Wäldchen & Mäder, 2018), can identify visual content on the legal and illegal wildlife trade. Supervised machine learning aims to learn how to map inputs (data) to outputs of interest by applying specific algorithms. In the complex case of bird trade, it would be helpful to train an algorithm that can process unseen data and label images with classes of interest (e.g., species).

Most research attention towards machine learning in conservation has been for species identification from camera trap images (Norouzzadeh et al., 2018; Tabak et al., 2018). Nevertheless, there have been significant leaps forward in image-based classification methods in conservation science. This progress has been advanced by camera trapping initiatives and data repositories, such as Google Earth's Wildlife Insights project and Microsoft's AI for Earth program. Species identification tools in the field of wildlife trade have lagged compared to these.

Machine learning tools such as CNN's, are inspired by the mammalian visual cortex (Norouzzadeh et al., 2018; Wäldchen & Mäder, 2018). Thus, they are valuable architectures to compare with manual, human visual identification of species, such as birds. Automatic visual recognition algorithms have recently achieved human expert performance at visual classification tasks in various disciplines from field biology to medicine (Beery et al., 2018). However, whether the same algorithms can achieve or outperform human expert performance in identifying birds in the wildlife trade remains unknown.

1.1 Aims and objectives

Although there has been a general acceleration in the number of artificial intelligence solutions in conservation, we need to have an accurate baseline of human ability to determine how well different algorithms perform and how computer algorithms perform against humans in identification tasks. Given the magnitude of trade in Indonesia and China, we will focus on birds traded in these countries.

Firstly, a repository of photos for birds that appear in the wildlife trade needs to be assembled. Having a photo library is critical to investigate the problems around bird identification further. Secondly, baselines of human identification error when classifying birds in wildlife markets are urgently needed, especially in the face of the ongoing Asian Songbird Extinction Crisis. Given the complexity and messiness of photo data from wildlife markets, proof-of-concept for the applicability of image-based machine learning approaches is also needed.

In the first data chapter (Chapter 2), we explore common facets of human performance in identifying animals and introduce the identification of birds and the theoretical challenges for humans. To build a baseline of human error in identifying a subset of birds that appear in the wildlife trade, we use match-mismatch experiments to quantify identification error, confidence, and agreement amongst a non-expert cohort.

In the second and final data chapter (Chapter 3), we construct a novel computer-based machine learning model which can accurately detect 37 species of bird. We discuss the process of selecting the correct model for this unique data problem and how a dataset can be prepared and optimised to train a deep learning model.

In the final chapter (Chapter 4), we discuss these results and their implications. Lastly, this human baseline and the performance of the most accurate computer model are compared. The creation of new ‘human-in-the-loop’ monitoring protocols for the wildlife trade is explored, using birds as our study taxon. We detail the value of our application and its future modifications to monitor legal bird trade, detect species bordering on extinction and expose wildlife crimes.

Chapter 2 Creating a baseline of human-based identification in the wildlife trade

2.1 Introduction

We must first identify species to conserve them, thus accurate identification is crucial (Farnsworth et al., 2013). Beyond conservation, identifications down to the rank of species are also needed for effective action on biosecurity and crime (Stringham et al., 2021). Traditionally, identifying animals involves a detailed visual inspection of the individual and comparison with reference images or specimens (Rahimi et al., 2016).

If identification is not performed to species level or species are misclassified, there are several risks. Conn et al. (2013) describe three main types of error in species identification; 1. misclassification, 2. partial observation, and 3. misclassification *and* partial observation. Misclassification is where participants assign a species to every observation, even if they are unsure; partial observation is where participants assign a species; if they are unsure, it is listed as unknown. Lastly, in misclassification *and* partial observation – species misclassification is combined with a new class for unknown species.

In recent years, more literature has emerged which measure's human ability in the identification of animals. Austen et al. (2016, 2018) showed that species identification could be difficult both for the layman and expert contributors to community science projects. Verifying species identities is further complicated when images are of low resolution or only show parts of an animal (i.e., missing key characters (Austen et al., 2016; Gibbon et al., 2015). For example, Gibbon et al. (2015) used match-mismatch experiments, a methodology commonly used in psychology, to assess individual identification accuracy of the mountain bongo, *Tragelaphus eurycerus isaaci*) amongst experts and non-experts, using images showing different traits of the bongo.

Species misclassification or identification errors can be elevated depending on the surveying method. Species misclassifications can be high for surveys that rely on auditory cues for observations made from a distance, such as a frog call survey or a land bird point count (Conn et al., 2013). Misclassifications can be high from images. Gooliaff and Hodges (2018) assert that even low misclassification rates can lead to the over or underestimating species distribution or habitat preferences. Johansson et al. (2020) showed that identification errors by experts led to the overestimation (on average by one-third) of snow leopard population abundance from camera trap surveys.

It is crucial to explore the potential impacts of identification errors of birds in wildlife markets, as these errors can be severe in applied settings (Alenezi et al., 2015). Unobserved declines in certain species (Gibson et al., 2019) sold at wildlife markets can result from false sightings of species or failures to detect the true market species diversity. Alfino and Roberts (2019) suggest that species misidentifications impact the accurate estimation of viable harvest levels for wildlife, as well as the detection of illicit trade. Further, suppose misidentification is high on the supply side of the trade chain. In that case, this could result in the accidental collection of non-target species due to errors in identifying or the deliberate laundering of target species as non-target ones (Alfino & Roberts, 2019).

The predominant methods for surveying wildlife markets are usually presence-absence surveys (a form of species classification). Here, a species list is created, or both the species identity and abundance are recorded by counting openly displayed individuals offered for sale at individual shops, using the widely practised Direct Counting Method (DCM) (Bušina et al., 2020). In theory, humans are adept at navigating through occlusion to recognise objects (Chandler & Mingolla, 2016). In wildlife markets, these occlusions may be cage bars, other people (traders or customers at the market), or other birds in the same cage, which together or separately may obstruct a clear view of a bird being offered for sale.

Given the ongoing Asian Songbird Crisis, the existence of identification errors when surveying markets is of serious conservation concern. Whilst trained experts in Southeast Asian bird taxonomy do exist, they hold the knowledge that is not common in a population and are hence rare. Experts such as taxonomists are unlikely to be present on routine surveys. As a result, large-scale surveys may be infrequent due to their availability (in addition to other time and financial constraints). For example, Chng et al. (2015), along with a team highly experienced in Southeast Asian bird taxonomy, counted 206 species across three markets in Jakarta during a 3-day window.

Although this is an immense number of species, the total count may have been higher since the surveyors could not identify every individual to species level. These species were subsequently excluded from the analysis. Thus, even if experts are the ones to survey markets using the partial observation method (Gooliaff & Hodges, 2018), this leads to the underestimation of both species richness and volume at markets.

The Hill myna (*Gracupica* spp.) and white-eye (*Zosterops* spp.) complexes are usually included in this 'unknown' category and are challenging to identify in markets (Chng & Eaton, 2016; Indraswari et al., 2020; Rentschlar et al., 2018). These groups experience high levels of cryptic diversity, and in particular, the white-eyes have experienced rapid speciation (Lim et al., 2019). Families that experience cryptic diversity frequently undergo taxonomic changes, which further complicates their identification as they can be misnamed. Alfino and Roberts (2019) also point out that the diversity in cryptic species (in their case, chameleons) makes identification challenging, particularly for non-specialists such as customs officers. Given the risks of misidentification errors of birds in wildlife markets and the highly speciose nature of the trade in Southeast Asia, we need to establish what human participants struggle with when identifying birds in wildlife markets before we approach individual species identification.

A greater understanding of wildlife markets is entrenched in the correct identification of species. Within the psychology literature, many studies have used untrained students as proxies for law enforcement personnel. This current study will thus provide a baseline of human ability to distinguish between species and quantify agreement and self-confidence of students in a match-mismatch task for birds in the wildlife trade. This chapter will also outline a methodology for generating human baselines for misidentification errors in the wildlife trade, which other researchers can easily apply to other taxa.

2.2 Methods

2.2.1 Ethics

The School of Anthropology and Conservation ethics committee at the University of Kent approved this research in August 2020 for initial data collection and December 2020 to launch the match-mismatch survey.

2.2.2 Data collection

Given the continuation of lockdowns in the UK, it was not possible to visit markets for research purposes between 2020 and 2021. However, many photos are available from various trade scenarios, including online trade, physical markets, and seizures.

We assembled a small dataset of labelled pictures (ca. 3,000 images) containing birds, taken across different retail and animal husbandry settings, for 19 species. The complete species list can be found in Table A1.1, hereafter referred to as the "MA_MIS_19" dataset. Sources include wildlife markets, mobile vendors, conservation breeding centres, zoos, and public groups on a large social media website. A range of different data providers was sought, including e-commerce sites, wildlife-trade-focussed NGOs, and large zoo consortiums. On social media, enthusiast groups primarily operate in Bahasa Indonesian. They were found using the keywords 'burung' and 'kicau mania' (songbird craze), followed by key locations in Indonesia that are implicated in the bird trade. Another portion of the images was downloaded from search sites (the open web) such as Google when searching for the English species name, plus keywords such as cage, caged, pet, and trade. Submissions were accepted via a designated Gmail account.

2.2.3 Participant recruitment

Since untrained students are commonly used as participants in psychological experiments in this field, we recruited students from the University of Kent to participate in the match-mismatch experiment, with the incentive of winning a prize upon completion. This cohort is a proxy for non-specialists such as customs officers (Alfino & Roberts, 2019). All the participants were master's students in conservation biology or postgraduate researchers within the Durrell Institute of Conservation and Ecology in the School of Anthropology and Conservation. All participants gave prior informed consent.

2.2.4 Survey design

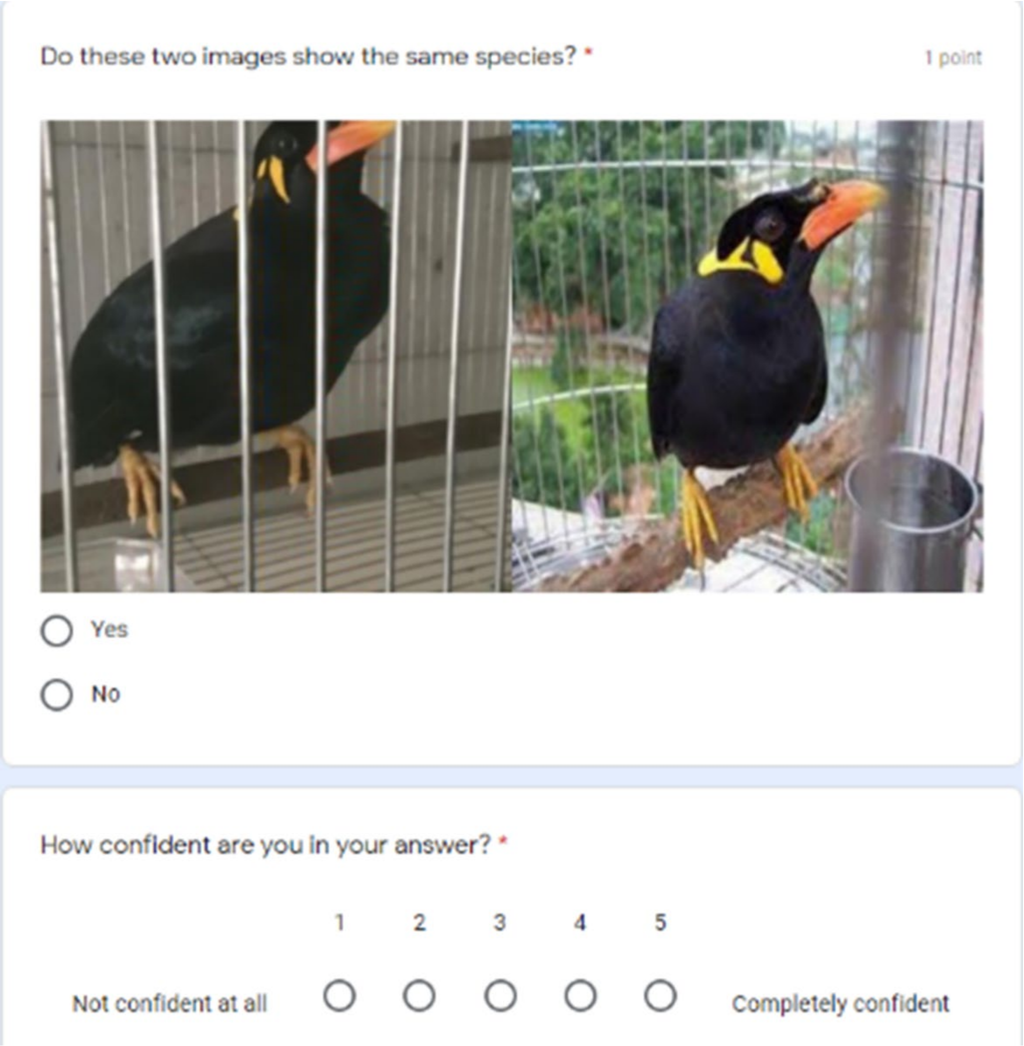
Unlike comparable studies using match-mismatch experiments, such as Alfino and Roberts (2019), we were interested in developing a random sample analogous to what a computer may view in training, where training samples are shuffled and shown in batches for a neural network to learn. All data were pooled, then 400 images were randomly selected. These images were resized, then randomly stitched together to create side-by-side image pairings of matches and mismatches. Randomising the stitching eliminated the possibility of selecting photos that were too similar or dissimilar. If we had manually chosen images, this might not have been a fair test of distinguishing species apart. In addition, given the variation in size, age, and condition of birds in wildlife markets, randomised simulations of this kind are closer to the real-life situation of a wildlife market. This generation of randomised stimuli was performed in Python, version 3.8 (code available in this Github repository: https://github.com/Sicily-F/cagedbirdID).

Each image pairing was verified a priori as a match or mismatch pair, allowing us to assess whether the classifications by participants were correct. During this verification process, the pairings were re-checked so that no image was shown twice in the survey to reduce the possibility that participants were not learning the species identity through doing the study itself. It was estimated that each participant required 5 seconds per pairing, and thus approximately 20-25 minutes to complete the task of 200 image pairs.

The survey was active from December 2020 to January 2021, hosted on Google Forms, as the results can be exported in .csv format. The test consisted of 200 pairs of images of birds, following our random merging of these pairs, 5% were species pairs, and 95% were pairs of different species.

In each question, the participants were prompted to determine if the species in each image were the same or not, “yes” (match) or “no” (mismatch). This is known as the two-alternative forced-choice (2AFC) paradigm (employed in related work by Gibbon et al., 2015), where no option for ‘unknown’ or ‘not sure’ was provided. The order of pairings in the task was random but was the same for all participants. The participants did not know who else was participating in the experiment.

Like Chizinski, Martin, and Pope (2014), who assessed self-confidence amongst anglers in identifying angler fish, the participants were asked to rate their self-confidence for each question on a five-point Likert scale. In this scale, 1 = not confident, 2 = a little confident, 3 = somewhat confident, 4 = very confident and 5 = extremely confident. Each score on the Likert scale was converted to a percentage as a measure of confidence. For example, if a participant described themselves as very confident on the Likert scale, this was treated as a fraction of 4/5 (0.8).



Do these two images show the same species? * 1 point

○ Yes

○ No

How confident are you in your answer? *

	1	2	3	4	5	
Not confident at all	○	○	○	○	○	Completely confident

Figure 2.1: An example of a match side-by-side pairing, in the Google Forms survey. Each question was the same, with the Likert scale observed underneath.

2.2.5 Assumptions

Before the survey, we assumed that each participant did not have detailed knowledge of the taxonomy of Southeast Asian bird species which appeared in trade. This is akin to law enforcement officers who may survey markets or are involved in units who perform seizures before any specific training occurs. Our participants, therefore, represent a non-specialist, educated cohort who would likely be able to distinguish between images of birds. When this study was performed, there were no images of species included in the stimuli that were sexually dimorphic. Thus, in each answer, the observer was instructed to assume that the species are not sexually dimorphic and that if individuals looked different, they were likely different species. It was also assumed that each observer completed the task in one sitting.

2.2.6 Calculating metrics of performance

The precision, recall, and F1 score (the harmonic mean) for the participants was calculated. The precision is the proportion of correct positive identifications, the recall is the proportion of actual positives identified correctly, and the F1 score is the harmonic mean of both metrics. They are calculated based on the true positive, the true negative, the false positive, and the false negatives rates (explained in Table 2.1).

Table 2.1: An explanation of the terms true positive, the true negative, the false positive, and the false negatives rates in the context of our experiment.

	Expert-identified species (actual)		
Participant identified species (predicted)		Match	Mismatch
	Match	*True positive* The participant identified the pairing as a match, and it had been pre-determined as a match.	*False positive* The participant identified the pairing as a match, though it was a mismatch.
	Mismatch	*False negative* The participant identified the pairing as a mismatch, but it was a match.	*True negative* The participant identified the pairing as a mismatch, and we pre-determined that it was a mismatch.

The following equations describe how precision, recall, and the F1 score are calculated.

$$Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$F1 = 2 \cdot \frac{\text{precision} * \text{recall}}{\text{recision} + \text{recall}}$$

The proportion of agreement for match and mismatch questions was also calculated amongst participants, following the equation used by Gooliaff and Hodges (2018). A modified version is provided here, where *Match* and *Mismatch* are the numbers of participants that classified an image as “match” or “mismatch” respectively, and *n* is the total number of non-experts:

$$Agreement = \frac{[(match^2 + mismatch^2) - n]}{n \text{ x } (n-1)}$$

With two classification options, the proportion of agreement had an upper bound of 1.00 (perfect agreement) and had a lower bound of 0.48 (near-perfect disagreement). In this case, this would be how many participants classified an image using the incorrect option out of “match” or ”mismatch.”

2.2.7 Statistical analysis

The data is non-parametric since the sample size of match and mismatch questions were not equal. Five percent of the questions were match questions, and 95% were mismatch questions. The Wilcoxon rank-sum test with continuity correction was performed to test whether there was any difference between the mean of the match and mismatch scores. This test is used to compare measures of location when the underlying distributions are not normal or known in advance (Rosner et al., 2003). It reveals whether the mean of one group is significantly different from another group.

The F-test is commonly used to compare the two standard deviations of two samples and check the variability. Again, a non-parametric alternative was used to test the ratio of the two estimates of variance, match, and mismatch questions. The Ansari-Bradley test returns a test decision for the null hypothesis that the data in vectors x (match questions) and y (mismatch questions) come from the same distribution. The alternative hypothesis is that x and y come from distributions with the same median and shape but different dispersions (variances). We also tested for any correlation between the metrics for accuracy, confidence, and agreement. We calculated Pearson's correlation coefficient to test for correlation between metrics.

Law enforcement personnel who have excelled in facial recognition studies have been named 'super-recognisers' (Robertson et al., 2016). We sought to detect the presence of super-recognisers amongst our participants by using the Hampel filtering approach to detect any outliers (Verma et al., 2019). Outliers are data points that differ significantly from other observations.

The Hampel filter identifies outliers based on certain thresholds from the median values (Verma et al., 2019). In this case, an interval (I) of plus or minus three median absolute deviations (MAD) was used. The MAD is the median of absolute deviations from the median for all the data. The equation for the interval to identify outliers is below:

$$I = [median - 3 \cdot MAD;\ median + 3 \cdot MAD]$$

If the data distribution is normal when the outliers were removed, a statistical test called the Dixon test can be used. The Dixon test can be used to see if the highest or lowest values in a distribution are significant as an outlier with respect to the whole distribution. The Dixon test is helpful for small sample sizes, which were expected in the short time frame in which our survey was live. We searched for a high outlier (i.e., a super-recogniser) in our dataset, using the Dixon test in the 'outliers' package in the R statistical program (Komsta, 2011).

As is good practice, in addition to Hampel filtering and statistical tests, a boxplot was also plotted to display all potential outliers.

2.2.8 Phylogenetic analysis

We calculated the average accuracy (number of correct answers) across all questions involving each species. For example, if the Chestnut-backed Thrush (*Zoothera dohertyi*) appeared in 10 questions, once with itself and on nine occasions with different species, we averaged the accuracy for all these questions.

To investigate if accuracy is similar amongst related species, we plotted a phylogeny of the 19 species plotted, along with their associated accuracy. A phylogenetic subset of the species in the MA_MIS_19 dataset was downloaded from the BirdTree project (available at www.birdtree.org). A 100-tree subset (the minimum) was selected and downloaded. We used a majority-rules consensus phylogeny to build a consensus tree. A majority rules consensus phylogeny includes only clades (a grouping that includes a common ancestor and all descendants) present in most of the trees in the original set. The phylogeny plotted had a different nomenclature to the complete species list (Table A1.1) since the BirdTree data uses an older taxonomy than the BirdLife taxonomy (data retrieved at: http://datazone.birdlife.org/home).

Using the ggfree package (Poon, 2019), the phylogeny was plotted and labelled with the accuracy for each question in which the species appeared. The range of the plot region is determined by the branch lengths of the tree itself. Hence, species that had the longest branch radiating out from the centre of the plot are the most evolutionarily distinct.

All statistical and phylogenetic analyses were conducted in R version 4.0.4.

2.3 Results

Twenty-seven participants completed the survey. Overall, the mean accuracy amongst our participants was 92.8% (±0.10 SD) (Table 2.2).

Table 2.2: Summary of different metrics relating to all questions, match questions only, and mismatch questions only, to 3 significant figures.

	All questions	Match questions	Mismatch questions
Mean accuracy (% correct answers)	93.8	74.4	93.7
Confidence	80.6	67.0	81.3
Agreement	77.6	56.1	78.7
True negatives			93.3
False negatives			25.6
True positives		74.4	
False positives		6.70	
Precision	91.7		
Recall	74.4		
F-score	82.2		

Participants were better at discriminating between species than they were at matching. Accuracy was higher for match questions (mean = 93.8%, ±0.09 SD) than for match questions (mean = 74.4%, ±0.16 SD). Further, this difference was highly significant (Wilcoxon = 1721, p-value = 0.826 e-06). In addition, the difference in variances between the two classes of questions was also significant (Ansari-Bradley test = 10091, p-value= 4.73 e-08). Both confidence (mean = 4.03, 80.1%, ±0.54 SD) and agreement (mean = 75.6%, ±0.11 SD) were also lower for match questions.

The F1 score of 82.2 is a truer measure of performance than the average accuracy across all questions (Table 2.2) since it also accounts for the true and false negatives rates.

Table 2.3: Correlation matrix between accuracy, confidence, and agreement across all questions, to 3 significant figures.

	Accuracy	**Confidence**	**Agreement**
Accuracy	1.00	0.564	0.834
Confidence	0.564	1.00	0.584
Agreement	0.834	0.584	1.00

There was no strong correlation between accuracy and confidence nor agreement and confidence (Table 2.2). However, there was a strong correlation between accuracy and agreement (Pearsons=0.84). In this case, the accuracy was higher in questions where most participants agreed.

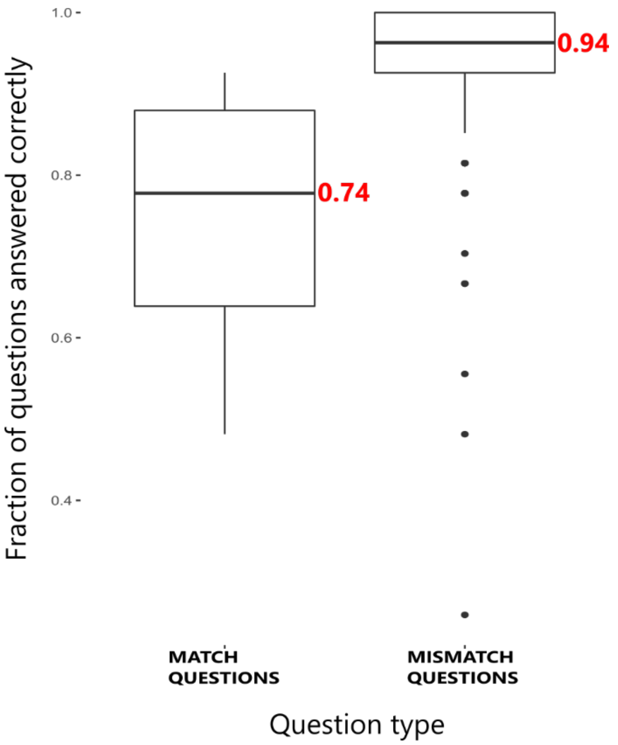


Figure 2.2: A boxplot of all participants' mean, standard deviation, and variance for the match versus mismatch questions.

The range of scores was greatest for mismatch questions (Figure 2.2). However, most scores for mismatch questions were in a more constrained range than for match questions.

2.3.1 Testing for outliers

The Hampel filtering approach provided an outlier range of 0.895 to 1.015, using the default threshold for the approach. It also revealed four outliers, all of which fell beyond the lower bound of 3 median absolute intervals below the median itself (Figure 2.3). The Dixon test result, to see if the top result (0.995) was an outlier, was non-significant (Q = 0.091, p-value = 0.313). The boxplot also showed the same four outliers, which can be seen in Figure 2.3 and were the four lowest scores.

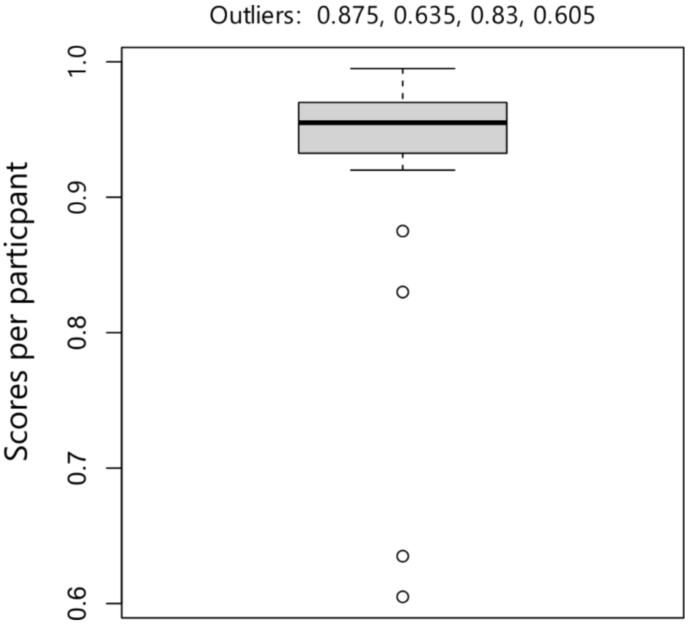


Figure 2.3: A boxplot of all participants' mean, standard deviance, and variance of scores across all questions. The four circles at the lower tail of the plot are the outlying scores.

2.3.2 Visualising the phylogeny

There appeared to be no clear trends of similar accuracies amongst related species (Figure 2.4). However, accuracy seems to be lower amongst some of the thrush family (Turdidae), specifically the Chestnut-capped Thrush (*Zoothera interpres*, an accuracy of 0.90) and Chestnut-backed Thrush (*Z. dohertyi,* an accuracy of 0.89). The lowest accuracy (0.83) was recorded for a close relative in another family, the Bluethroat (*Luscinia svecica*). The highest score was for the Greater Green Leafbird (*Chloropsis sonnerati*), with an accuracy of 0.97.

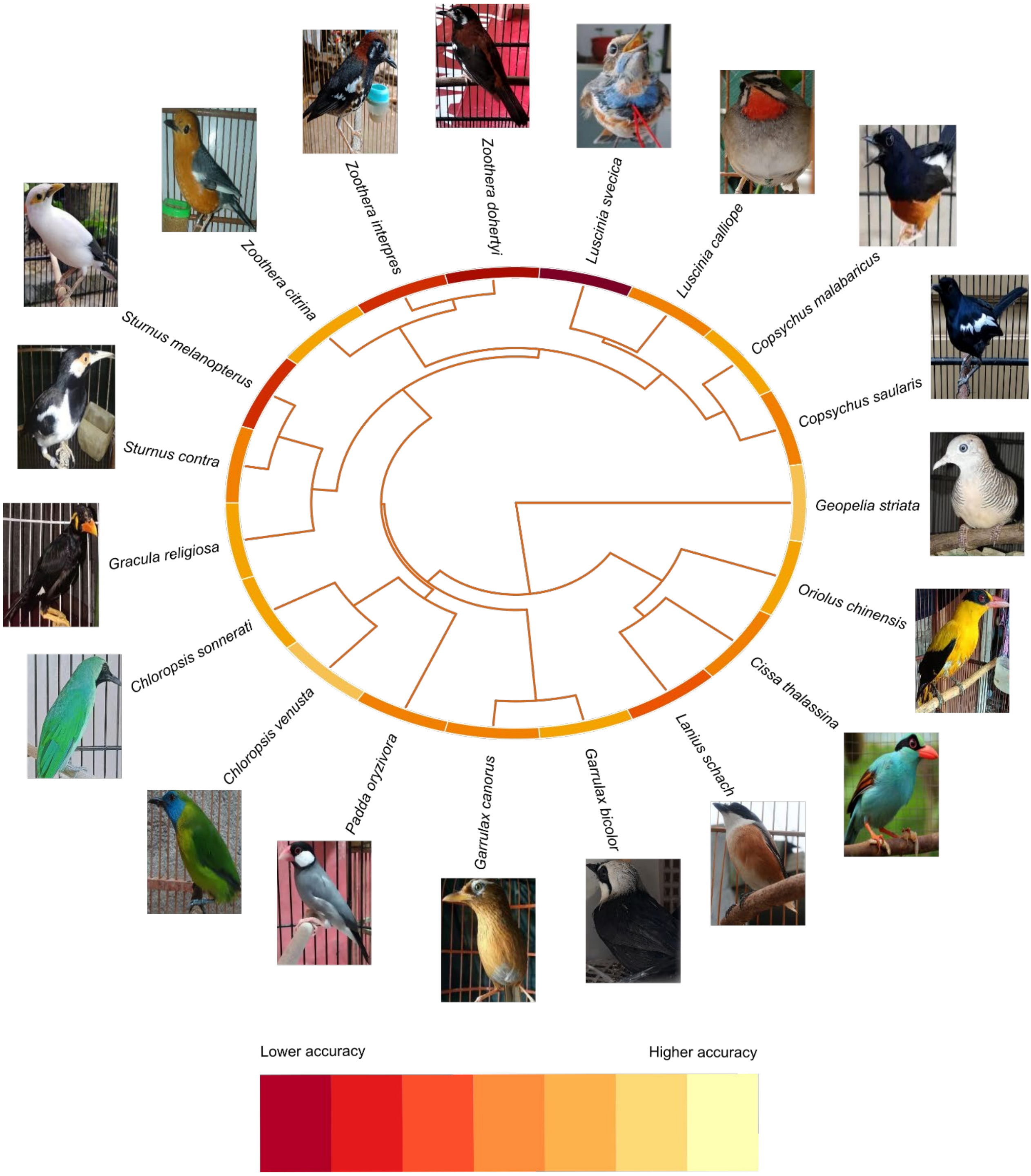


Figure 2.4: A radial phylogeny showing the evolutionary relationships between the nineteen species included in our match-mismatch experiment. Each species is marked with the averaged accuracy for all questions in which the species appeared. Lighter colours are those with higher accuracy. The legend is defined since the tree is centred on the origin (x=0, y=0). Hence the negative values are not a 'negative' branch length; they are just on the left of 0.

2.4 Discussion

This study provides a baseline of human misidentification error for birds that are popular in the bird trade in East and Southeast Asia. Given the ongoing risk to birds in Southeast Asia, the use of image recognition technology can aid the process of species identification. Before such technology is deployed, it needs to be measured against the effectiveness of humans in comparable tasks.

2.4.1 Match versus mismatch questions

Mismatch questions had higher overall accuracy and a higher true positive rate (Table 2.2). Thus, participants were better at mismatching tasks than matching tasks. This contradicts the results of similar studies for individual bongo identification (Gibbon et al., 2015) and chameleon species identification (Alfino & Roberts, 2019), where overall error rates were higher for mismatches than for matches. Although mismatch accuracy might seem initially high in our study, if more trials were performed, performance would likely decline on mismatch trials throughout the experiment (Alenezi et al., 2015). In an extensive face identification experiment (c.a. 1,000 trials), Alenezi et al. (2015) found that mismatch accuracy declined over time, whereas there was a concurrent increase in matching accuracy. Though the average accuracy was lower for match questions (Table 2.2, Figure 2.4), this may simply be an artefact of sample size, as there were only ten match questions compared with 190 mismatch questions.

The relatively high error rate for match questions (74.4%, an error rate of 26.6%) could be because our participants were unfamiliar with bird identification and had never visited a bird market. A high error rate for matches may be partially explained by the use of randomised images, resulting in some blurry, dark, or partially occluded images being presented to participants. Our error rate is much higher than the match error rate recorded by Alfino and Roberts (2019) (4.8%). However, our result is consistent with bird market studies by the likes of Chng et al., 2015; Indraswari et al., 2020; Rentschlar et al., 2018, who assert that similar species are challenging to identify.

Alfino and Roberts (2019) recorded 14.3% for their mismatch error rate, whereas we observed a much lower error rate for mismatches (6.4%). However, it is important to note that Alfino and

Roberts (2019) looked at identification within a single genus. This contrasts with our study, a highly diverse group where differences may be more evident than in certain taxonomic complexes of birds such as the genus *Zosterops*.

However, the fact that our participants were not trained may have no bearing on results anyhow. Occupational experience (i.e. years of service) does not necessarily improve matching accuracy (Robertson et al., 2016). Johansson et al. (2020), in their study on the individual matching of snow leopards amongst experts, showed that experienced participants make mistakes and that these mistakes were reasonably common (5% of observations were misclassified). However, individual identification and even species identification is arguably a much more challenging task than matching task for species distinction.

Despite mismatch questions being presumably easier for participants, even for mismatch questions, there is considerable variability in accuracy amongst participants (Figure 2.2). In their similar study on look-alike species in a genus of Malagasy chameleons, Alfino and Roberts (2019) also found that identification error rates varied widely, reaching high error levels for 'look-alike species'. In the present study, the closely related pair, the Chestnut-capped Thrush (*Zoothera interpres*) and the Chestnut-backed Thrush (*Z. dohertyi*), also had high error rates (low accuracy). Although our accuracy is computed differently from their study, Alfino and Roberts (2019) defined low identification error rates as those below 25%, which was the case for all our species.

2.4.2 Confidence levels

The self-confidence amongst participants for matches (67.0%) was lower than for mismatches (81.3%) (Table 2.2). In their study on angler fish identification, Chizinski, Martin, and Pope (2014) found that self-confidence increased with skill level. Eighty-one percent for mismatch questions is equivalent to giving four on the Likert scale, and 67% is akin to giving three or four (Table 2.2). Black et al. (2013), in a study on barbary lion extinction, asked experts to score sources of evidence on lion sightings. Generally, the experts all gave the evidence an 80% in terms of reliability, which is expect given that four out of five is known to be the 'lazy option' on surveys. Thus, this self-confidence result may be an artefact of human behaviour rather than a lack of confidence in decision-making during the experiment. Nevertheless, this evidence suggests that, amongst our participants, although they know how to distinguish between species (overall accuracy of 93.8% correctly identifying matches and not matches), they are only mildly confident in their ability to do so (Table 2.2).

The result of relatively low confidence may be more associated with the images' qualities, such as blur, low resolution, and partial image capture, which can complicate identification even for experts (Gomez-Villa et al., 2016). Alenezi et al. (2015) suggested that changes in lighting, expression, or view can induce many differences in the appearance of a face; the same could be true for birds. Alternatively, body size could play a role; Gibson et al. (2019) found that smaller sharks were harder for anglers to identify. Alternatively, it may be more straightforward that the species themselves were unfamiliar to the participants; therefore, they could get questions right but were not necessarily very confident in their decision.

This ‘underconfident’ behaviour contrasts with the behaviour observed by Chizinski, Martin, and Pope (2014), who found that anglers may be relatively confident in their ability to identify a species. However, our participants still may not know how to distinguish a particular species, and our results would suggest that bird species identification could be perceived as more difficult than other tasks. That said, there are many interventions which can be used to boost the confidence of observers, none of which were employed in our specific study. Chizinski, Martin, and Pope (2014) suggested that educational programs should target lower-skilled anglers due to a lack of self-confidence to identify species individually. If our participants had training, this might increase their confidence. It should be noted that Chizinski, Martin, and Pope (2014) examined a group of participants who had a specific interest in angler fish. If Southeast Asian bird enthusiasts were to be studied here, their confidence might have been higher.

Conversely, theory-based learning may not be the most effective means to achieve accurate species identification. Austen et al. (2016) found that the accuracy of bumblebee identification was higher and more consistent among professionals with field expertise than those with expertise gleaned from books. Hence, educational programs grounded in theory-based learning might not be very useful for birds either. Interestingly, Chizinski, Martin, and Pope (2014) also showed that 55% of the anglers with average skill still carried guidebooks. Along with those of average ability, 36% of those with the highest skill level still took a guidebook. Even those highly skilled in identification still benefit from visual aids, though they were less likely to carry a guidebook. The role of visual identification aids as a potential contributor to confidence levels should be explored further in the field of bird identification.

2.4.3 Agreement levels

Participants had a far greater agreement for mismatch pairings rather than match pairings. Disagreement was almost 'perfect' for match questions (Table 2.2, 56.1%). That said, the agreement was by no means perfect for match questions (Table 2.2, 78.7%). The reason for this relatively moderate agreement may be since all our participants were new to the study area. This lack of consensus could translate into a real-life situation, where customs officers or inspectors of wildlife markets could disagree on species identifications. This could prove problematic since usually more than one observer is tasked with surveying markets, and a lack of agreement may result in inconclusive species identifications. Yet, even in an identification task for highly similar, congeneric bobcat and lynx, Gooliaff and Hodges (2018) found that experts were inconsistent with their original identification in repeat trials.

These results show that non-specialists may struggle to recognise when matches occur and will have a lower agreement amongst each other and confidence in their classification (Table 2.2.). Gooliaff and Hodges (2018), in their study on expert's ability to distinguish between the bobcat, *Lynx rufus*, and the Canada lynx, *Lynx canadensis* found lower agreement (K =0.64) amongst experts than observed here, on a distinguishing task between two classes, and a further category for unknown pairings. Johansson (2020) found that experts made one-third fewer classification errors than non-experts on a task for individual snow leopard recognition. For the task of Southeast Asian bird matching and mismatching, we would expect accuracy, agreement, and confidence to be higher amongst experts than that reported here. Alfino and Roberts (2019) quantified demographic aspects of participants, including their vision (good or corrected-to-normal). In future work, we could record a variable for vision that might further explain our participants' variation.

2.4.4 The existence of super-recognisers

One observer achieved 99.5% on the matching task. The next closest score was 98.0%. Although these scores are close, considering the range of scores from our participants (60.5-99.5% - Figure 2.3), the fact that this participant only gave an incorrect answer to one question is highly impressive. In the field of face recognition, super-recognisers have been well documented (Robertson et al., 2016; Russell et al., 2009), though they represent a tiny fraction of the population. However, the results from the Dixon test showed that this high-scoring participant was not a significant outlier (p-value = 0.313), so they cannot be confidently defined as a super-recogniser. Conversely, it could also be the case that our low sample size limited the statistical power of the Dixon test. If we had recruited more participants, this individual's performance might have become an outlier. The Hampel filtering approach and visualisation of the accuracy (Figure 2.3) showed that four participants were below average and likely not suited for tasks of this kind. Amongst our 27 participants, this equates to 14.8 % of the cohort. In studies such as this, where participants have no occupational experience, it is still likely the case that some participants are less gifted at matching tasks, particularly for complicated cases such as identifying birds in the wildlife trade. The other participants may have scored so highly because the survey did not contain many closely related species (within the same genus). Therefore, the highest human score may be superficially high on a relatively simple task. Nonetheless, these results lay a foundation for a diagnostic framework for moderate and high performers in identification tasks, which may help to concentrate the capacity of law enforcement agencies based on the ability of their staff.

2.4.5 Phylogenetic pattern of accuracy

The phylogeny shows that accuracy is relatively heterogeneous amongst related species, with a range between 0.83 and 0.97 of accuracies recorded (Figure 2.4). There was a range of 0.03-0.17 in accuracy amongst the species. Accuracies might be more homogenous if more sympatric species were involved. Our lowest accuracy was recorded for the Bluethroat, *Luscinia svecica* (83% accuracy) (Figure 2.4). However, this only appeared in seven questions in the whole survey.

Although no correlation was found between accuracy and the number of questions a species appeared in, at least for the Bluethroat, this may be explained by the fact that it can have higher false-positive and false-negative errors as a rarer species than common species (Gooliaff & Hodges, 2018).

2.4.6 Limitations

Class imbalance

One drawback of utilising the randomisation technique for generating the side-by-side pairings is that class imbalance was present in the MA_MIS_19 dataset. Therefore, some species appeared in more questions in our survey as they were better represented in the MA_MIS_19 dataset. Nevertheless, our method reproduces how a computer model would randomly receive photos to learn their features. It also represents real-life situations in wildlife markets, where some species are more abundant than others, known as 'wildlife hot products' (Moreto & Lemieux, 2015). On the other hand, for rarer species, some are only recorded once or twice in surveys. Hence market surveyors need to be skilled enough to make species identification upon only seeing a species once.

The time taken to complete the survey

Although it was estimated for each participant to take five seconds per question, some of the participants commented, post-completion of the task, that it had taken them longer than the described time. Nonetheless, so that a surveyor does not look suspicious or is mistaken for a customer, there may not be much time available to make an identification, so five seconds is realistic. Further, Google Forms does not have an internal function to record the time taken to complete a form or survey. Although we did not register the exact time taken, this may suggest that fatigue plays a role and that participants may need longer to make their choices as the survey progresses. Subsequent surveys of this nature should ask the participant to self-record the time taken to complete it.

Notwithstanding, as previously mentioned, in some markets there have been upwards of 200 species recorded in one city (Jakarta, Indonesia). Therefore, a participant in a wildlife market survey could be faced with hundreds of these decisions. Although in a different matching task, for human faces, Alenezi et al. (2015) found in an extensive face matching task (over 1,000 trials) that enforced rest and desk-switching cannot maintain identification accuracy. We hypothesise that fatigue and a continual decline in identification accuracy across many trials can result in lower quality data collected in wildlife markets that do not capture the species diversity or abundance present. If wildlife market surveys do not record accuracy, confidence, and agreement, as done here, we may unintentionally have poorer results as the survey progresses. Since these metrics are not usually published with survey data, we can only assume that some data of lower quality is inadvertently collected.

Since identification accuracy can degrade over time for extensive face matching tasks, this may be true of extensive species matching tasks for birds or other taxa involved in the wildlife trade. It is worth noting that this decline in identification accuracy is recorded even in quiet environments, where participants are alone, taking rests and switching locations (Alenezi et al., 2015). This decline may be even more exaggerated in the commotion of a wildlife market. On the other hand, this fatigue may be partially created by the monotony of facing a screen for long periods. Even so, rapid app-based identification methods may prove a time-saving method, as field expertise takes a long time to acquire. Training may not necessarily be beneficial to identify species with high accuracy.

2.5 Conclusion

This study demonstrates the importance of creating baselines of identification errors before automated identification is attempted. We have shown that the mean accuracy in distinguishing between species is 93.7% in non-expert observers. The likelihood of having participants who perform poorly may be higher, especially amongst an untrained cohort. Our data support the subsequent work in Chapter 3 to develop an automated species identification system. The data also provide a baseline of human performance against which the identification accuracy of said systems can be performed.

From this body of work, most importantly, as others have before (e.g. Alfino and Roberts, 2019), we have built an understanding of the patterns of identification error, which provides a baseline for the challenges faced by humans (e.g. law enforcement) in identifying birds. By first understanding patterns of accuracy amongst humans, we can develop software solutions against this established benchmark. These solutions are especially critical for the concentrated effort required to carry out market surveys. Many software solutions are advertised as time-saving. However, if they do not demonstrate improved accuracy upon a baseline of human performance, it may not be worth implementing such solutions.

Chapter 3 Developing a supervised, multiclass image classification model to identify birds present in the wildlife trade

3.1 Introduction

Machine learning has often been touted as a solution to monitor the highly diverse nature of several wildlife trades (Hernandez-Castro & Roberts, 2015; B. M. Marshall et al., 2020; Stringham et al., 2021; Xu et al., 2019). However, few practical and applied examples have tested the feasibility of the complex image data available in the wildlife trade. Photos can be complex for a variety of reasons, such as dimension, resolution, and blur. Still, image classification is of particular interest to monitor physical and virtual marketplaces for wildlife, though it is a relatively unexplored area of research. Most machine learning applications in the conservation field have focussed mainly on camera traps as sources for images (Olden et al., 2008; Schneider, 2019; Wäldchen & Mäder, 2018).

The general purpose of an image-based machine learning model is to learn how to classify categories (from a pre-labelled training set) and assign labels to images in an unseen test dataset. Pre-trained models are available (Choe et al., 2020; Das & Kumar, 2018; Weinstein, 2018), which have been trained on the ImageNet dataset. The ImageNet dataset has over 1,000 classes and 1 million images (Deng et al., 2009). In recent years, image classification has been dominated by large, open-source convolutional neural networks (CNN's), as demonstrated in recent large-scale visual recognition challenges (Willi et al., 2019).

CNN's can learn from images by directly analysing their pixels (Couret et al., 2020). CNN's generally consist of stacking groups of convolutional layers and pooling layers, inspired by biological visual systems (Beery et al., 2019; Christin et al., 2019). Transfer learning allows users to download pre-trained models trained on benchmark datasets such as ImageNet, with 1000 classes. Low-level feature representations, such as curves, shapes, and outlines of objects, have already been learned by algorithms (Choe et al., 2020; Couret et al., 2020). Since building models from scratch can be a slow process, transfer learning is a popular, computationally cheap method to retrain pre-built models on new image classes (Choe et al., 2020; Weinstein, 2018).

In general, classifying images is difficult when they are blurry, taken in poor lighting, and only show part of the animal due to occlusion (Gooliaff & Hodges, 2018). In the context of wildlife markets there are three main challenges when classifying birds from photos. These are issues with the 1. hardware or recording device of choice, 2. the location where the image is taken, and 3. the subject itself.

1. Hardware challenges

Using different hardware such as phones and cameras can result in blurred images from too short or long a time between frames and an inappropriate flash power for the photo's setting (Gomez-Villa et al., 2017). Although they do not hide the features of an animal (Gomez-Villa et al., 2016), blurred images can reduce the confidence of a species classification by altering its shape and colouration (Gomez-Villa et al., 2017). Overexposed images result when the animal is too near the camera's flash, and the external increase in illumination complicates classification (Weinstein, 2018). An increase in illumination can erase distinctive skin patterns or other animal features (Gomez-Villa et al., 2016). If the market is an indoor 'brick-and-mortar' one (Siriwat & Nijman, 2020), there may be a lack of illumination. For example, the iridescence of certain sunbird species can be hard to capture in poorly lit environments, so an accurate species identification can be challenging to achieve (James Eaton, October 2020, pers. comm.). Low-resolution images taken on lower quality cameras can be indistinguishable even for a human expert (Gomez-Villa et al., 2016; B. M. Marshall et al., 2020).

If data is collected using a video camera, issues of interlacing may arise. Interlacing relates to video scanning, where each frame's odd and horizontal lines are drawn on alternating passes. Interlacing transmits a full frame quickly to reduce flicker. Video frames can be deinterlaced to fill in the gaps between fields to create individual frames. Sibley et al. (2006) warned against the risks of interlacing, where motion is so quick that it causes noticeable differences in the positions of the fields. In the contentious case of the rediscovered Ivory-billed Woodpecker, *Campephilus principalis*, in Arkansas, USA, the evidence largely rested on video footage (Sibley et al., 2006). In their re-analysis of the footage, post-deinterlacing, Sibley et al. (2006) found features that supported the identification as a Pileated Woodpecker, *Dryocopus pileatus*, which contradict the initial identification as an Ivory-billed Woodpecker.

2. Location challenges

Th*e* location of where a bird is sold can also complicate photo classification. A common problem with nature photographs is that species identities cannot always be ascertained from a single image, particularly when similar species occur in the same locality (B. M. Marshall et al., 2020). Moreover, some congeneric species can only be told apart based on the location. For example, the male Javan Leafbird, *Chloropsis cochinchinensis*, found only on Java, looks very like the female Blue-winged Leafbird, *C. moluccensis* (distributed as far east as Borneo and south as Sumatra). If data for the sale location is available with each photo, then species identification is more feasible.

More specific than the sale location (i.e., a market in Jakarta), where the bird is photographed is also pertinent (i.e., in a cage in a market or a crate during a seizure). In general, classifying images is difficult when only part of the animal is in the picture (Gooliaff & Hodges, 2018), which may be due to occlusion. Occlusions are the pixels that obscure a clear view of a target object (Chandler & Mingolla, 2016). Occlusion can introduce noise into animal body regions in images (Gomez-Villa et al., 2016). In general, heavily occluded objects are more difficult to classify than obscured objects (Chandler & Mingolla, 2016). In wildlife markets, these occlusions may be cage bars, other people (traders or customers at the market), or other birds in the same cage, which together or separately may obstruct a clear view of a bird offered for sale.

3. The physical appearance of the bird

A central problem in species identification from photos is 'partial capture of an animal' (Gomez-Villa et al., 2016). Partial capture can be a problem in both manual and digital, computerised species identification. When only part of an animal is available to see, in real life or an image, certain features central to a species' identity are out of view.

The physical condition of the bird can also complicate their identification. Some wild species are very active in the cage, so they are hard to capture on camera; for example, the Blue-winged Leafbird, *Chloropsis cochinchinensis* is known to be very active in captive environments (James Eaton, October 2020, pers. comms.). In wildlife market environments, birds may be in poor health and have lost feathers or weight due to stress. Time spent in captivity, particularly on a poor diet, can also alter the visual appearance. Some insectivorous species sought after in the bird trade in Indonesia experience the 'blue-winged phenomenon'; namely the Javan Green Magpie, *Cissa thalassina* and the Greater Green Leafbird, *Chloropsis sonnerati*. The blue-winged phenomenon' is where the bird becomes a dull blue colour kept in captivity (Nijman et al., 2017) due to a lack of lutein pigment typically found in their insectivorous diet. For the sexually dimorphic Greater Green Leafbird, if females have been in captivity for a long time, they begin to look like female Blue-winged Leafbirds or Javan Leafbirds.

Male species are likely over-represented in the Indonesian bird trade context, given the current predilection for buying songbirds to train for singing competitions. Hence sexual dimorphism may not pose a massive challenge since females are rarer in specific trade sectors, resulting in a sexual skew. Beyond sexual dimorphism, there is inevitable variation in appearance within a species, including age (Gomez-Villa et al., 2017). Although adult birds are seen more at markets, juveniles and chicks are also present. In some cases, chicks may be more prevalent in the trade. For example, in bird markets in China, Red-whiskered Bulbul, *Pycnonotus jocosus*, chicks, a widespread urban bird, are frequently sold, either picked from nests or those which have fallen from the nest (Liang Zhijian, November 2020, pers. comms).

Although the Asian Songbird Crisis predominantly affects wild birds (H. Marshall et al., 2019; Sykes, 2017), consumers keep domesticated species for various reasons, with their song only being one. Several species are prevalent in captive-bred aviculture (Fischer's lovebird, *Agapornis fischeri*, the Grey Parrot, *Psittacus erithacus,* the Zebra Finch, *Taenopygia guttata*, and the White-rumped Munia, *Lonchura striata*). Certain colour morphs are popular in trade, which introduces artificial variation. For example, birdkeepers breed the White-rumped Munia to have unique colour patterns, which are appealing to some. Wild individuals usually have a stable colour, whereas the captive-bred ones tend to have a larger area of white colour (Liang Zhijian, November 2020, pers. comm.). Similarly, some wild-caught species are dyed different colours, such as the Red-billed Starling, *Sposiospar sericeus*, and the Scaly-breasted Munia, *Lonchura punctulata*, presumably to make them more attractive to prospective buyers (Figure A2.1).

3.1.1 Exploring the challenges of species identification for birds

Automated species identification is more difficult for large, homogenous groups (B. M. Marshall et al., 2020), such as the white-eyes (family Zosteropidae), all small-bodied and various shades of light green with white eyerings. Several methods have been proposed for species identification of birds, though primarily using audio data (e.g. Khalighifar et al., 2021). Nonetheless, automated image classification holds excellent potential for streamlining species identification (B. M. Marshall et al., 2020), particularly for birds.

In a particularly noticeable piece of work, Couret et al. (2020) applied different neural networks to distinguish 17 classes of mosquito genus', species and species networks and found that the DenseNet-201 network presented the highest validation accuracy at 96.20%. Elsewhere in the literature, there have been bird-specific applications of image-based machine learning using different architectures. Das and Kumar (2018) won the 2018 International Conference on Computer Vision & Image Processing by performing supervised classification of 16 species of wild birds found in the Himalayas using a dataset of 300 images. They performed multistage training with an ensemble containing the Inception V3 and Inception ResNetV2 architecture, though their highest training accuracy was 93.97% on cropped images using Inception V3. Niemi and Tanttu (2018) built a classifier for eight species encountered in offshore wind farms in Finland, using a Support Vector Machine classifier and achieved a reported sensitivity of 0.9463.

Choe, Choi, and Kim (2020) used transfer learning to identify four endangered parrot species from images taken in a zoo and downloaded from Google. They achieved an accuracy of 94.125% for the best performing model, a NASNetMobile model. In a study most similar to this, Ferreira et al. (2020) applied transfer learning for individual recognition of three bird species in wild and captive contexts, using the VGG-19 architecture to learn 30 classes of bird individuals. They achieved their highest training accuracy (92.4%) on individuals from one of the three species.

More recently, Fink et al. (2021) used a ResNet50 model to filter out relevant sales items containing birds from Facebook posts in Indonesian bird hobbyist groups. Without training the model further, they used existing ImageNet classes to predict the classification of posts. Posts either fell into a bird category, or where the cage was visually dominant, the categories of prison, shopping cart, or shopping basket.

In this research a methodology is presented for collecting data, localising birds in images, and training various machine learning models for species identification. First, open-source tools are applied to locate and crop 'messy' photos of birds in the wildlife trade. Then, we detail the stages for data pre-processing and the training of machine learning networks. We highlight two approaches for augmenting our training dataset. The first approach makes modifications to the original images, and the second generates artificial occlusion by superimposing pixellated boxes of random sizes onto images. Finally, we test the performance of the best model on datasets with differing percentages of occluded images.

We explore the following research questions:

1. Can transfer learning be used to distinguish between photos of bird species of varying quality and sources?

2. How does model performance vary when different networks and hyperparameters are tested?

3. Is model performance consistent when performing cross-fold validation?

4. Does the occlusion of birds in an image significantly affect the ability of the models to classify photos?

3.2 Methods

3.2.1 Ethics

We sought ethics approval to download photos from various online sources to build our database. The School of Anthropology and Conservation ethics committee at the University of Kent approved this research in August 2020 for initial data collection. The methodological pipeline for this study can be seen in Figure 3.1, detailing the data collection process to finally porting the model to a local application.

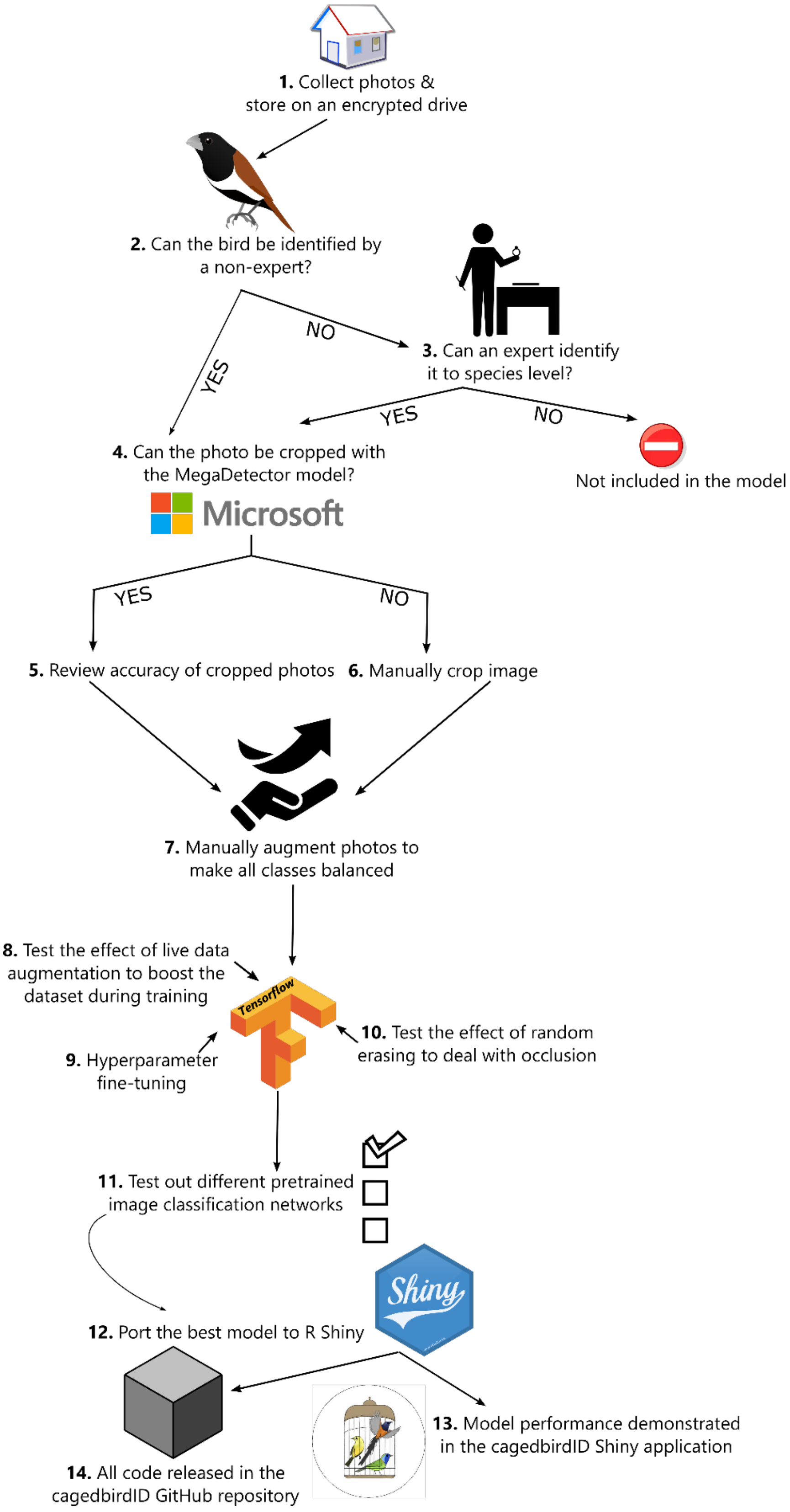

1. Collect photos & store on an encrypted drive
2. Can the bird be identified by a non-expert?
NO
YES
3. Can an expert identify it to species level?
YES
NO
Not included in the model
4. Can the photo be cropped with the MegaDetector model?
Microsoft
YES
NO
5. Review accuracy of cropped photos
6. Manually crop image
7. Manually augment photos to make all classes balanced
8. Test the effect of live data augmentation to boost the dataset during training
Tensorflow
9. Hyperparameter fine-tuning
10. Test the effect of random erasing to deal with occlusion
11. Test out different pretrained image classification networks
Shiny
12. Port the best model to R Shiny
13. Model performance demonstrated in the cagedbirdID Shiny application
14. All code released in the cagedbirdID GitHub repository

Figure 3.1: The methodological data collection and model training pipeline for image classification, created for this study.

3.2.2 Step 1: Data collection

It can be challenging to acquire sufficient training images of each species, especially for rare species. Building on the MA_MIS_19 dataset, we accumulated a dataset of labelled pictures containing birds (6,136 ground-truth images, 108-269 photos per species), taken across different retail and animal husbandry settings (see Table A2.1). In total, this resulted in 37 species in our complete dataset, hereafter referred to as the "TOT_SP_37" dataset. Locations include wildlife markets, conservation breeding centres, zoos, and public forums on a social media website. A range of data providers was sought, including e-commerce sites, wildlife-trade-focussed NGOs, and large zoo consortiums. Submissions were accepted via a designated Gmail account. The earliest recorded photo in the TOT_SP_37 dataset was from 2014, with the most recent photos in early 2021. The birds vary in terms of species, age, gender, and size, whereas the photos differed in dimensions, resolution, angle, and pose. For the most part, all images were of single, alive, adult, or juvenile individuals.

To gather data for popular species in Indonesia on social media, we searched public Bahasa Indonesian groups and used the keywords 'burung' and 'kicau mania' (songbird craze), followed by locations in Indonesia that are involved in the bird trade. When popular species were encountered, we saved this data. Another portion of the images was downloaded from Google when searching for the species name, plus keywords such as cage, caged, pet, and trade. A research assistant was also recruited to download photos from Baidu, the dominant internet search engine company in China. Two separate sites (Baidu Tupian, Baidu's specific image search engine; https://image.baidu.com/ and Baidu Tieba, Baidu's keyword-based community discussion forum; https://tieba.baidu.com/) were used to search for popular pet bird species in China, notably, the Siberian rubythroat, *Calliope calliope,* Bluethroat, *Luscina svecica*, Japanese Grosbeak, *Eophona personata*, Chinese Hwamei, *Garrulax canorus*, Red-billed Leiothrix, *Leiothrix lutea,* Silver-eared mesia, *Leiothrix argentarius* and the Zebra finch, *Taenopygia guttata*. These species were deliberately targeted as they are prevalent in China, easy to find open-source data for, and likely found in Chinese markets.

3.2.3 Step 2: Initial species identification and Step 3: Using expert verification

Our data providers often contributed photos with a pre-existing classification, either in folder form or file name. These classifications were then manually reviewed by the author. Where classification was not provided, these were manually classified by the author. If the primary author (S. Fiennes) could not verify the species identity or there were concerns about the reliability of the classification, another human classifier was consulted. The classification was lengthy for sympatric species, similar in size, shape, and colouration (Gooliaff & Hodges, 2018).

3.2.4 Data pre-processing

Raw photographic data obtained in wildlife markets is not suitable for input into a machine learning model. We first processed photos into a format that a neural network can receive as input. Given the wide range of data sources, photos were converted from various file extensions (e.g. .jfif, .png and .webp) to .jpg format.

The minimum number of ground-truth images considered for each species was c.a. 120 images. There are many reasons why data is either hard to collect or sparse; some species are not popular in the trade. Other species may be rare, so they only appear in markets sporadically. In addition, these may not be of high enough concern or interest to be housed in zoos or breeding centres, so photo opportunities are rare.

3.2.5 Step 4: Using the MegaDetector, an object detection model

Manually cropping photos is lengthy, especially since only one individual (the primary author) processed the data. A length cropping process could be a challenging bottleneck for this study, where some data submissions did not contain the bird in the foreground of the image, or there were many birds in one cage or image. There were also images where it was impossible to remove other birds from the background entirely. In TOT_SP_37, these were mainly species in the Estrildidae family, such as munias, finches, and weavers, which are sold to meet the intense demand for religious practices (i.e. prayer release) (Gilbert et al., 2012; Severinghaus & Chi, 1999).

Given the lack of open-source technology for image analysis of animals in the wildlife trade and other captive settings, we sought to test the feasibility of a camera-trap-oriented model to detect birds in images and crop them.

The MegaDetector model is typically used to detect animals and vehicles in camera trap images (Beery et al., 2019). Using the MegaDetector to crop images (around a bounding box) dramatically simplifies training species classifiers because images can be cropped to individual animals. Thus, our subsequent classifier has primarily animal pixels to learn from, rather than background pixels (Beery et al., 2019).

We tested the MegaDetector (version 4.1, 2020.04.27) from Microsoft AI to automatically locate and detect the birds in photos. This model is trained in TensorFlow on NC12v3 DSVM (Microsoft's cloud computing platform), Faster R-CNN (an object detection algorithm) with an InceptionResnetV2 CNN base, and pre-trained on the Microsoft COCO dataset (model and instructions available for download at https://bit.ly/3zAQ7OT). TensorFlow is a popular deep learning framework that uses a computational graph algorithm to represent mathematical operations as nodes and store data in multidimensional arrays (Weinstein, 2018). These algorithms can be used without further training by redeploying a 'frozen' computational graph.

The frozen, pre-trained MegaDetector was applied to the data, using the default confidence threshold of 0.85. In this case, the model is 85% certain that inside the bounding box is either an animal, person, or vehicle. Even if the model incorrectly classifies a bird as a human, this is of minimal concern because it still crops the 'living' object in the photo. Then, each image was cropped around the bounding box and outputted to a folder locally.

3.2.6 Step 5: Reviewing the MegaDetector output and 6. Cropping photos manually

Each cropped photo was then manually reviewed for input into our model. If the MegaDetector was not successful in cropping, we manually cropped images. Although some of our images incidentally contain people in the background, we also removed these using the MegaDetector, ensuring an ethical training process, which only included avian subjects and minimised 'human bycatch' (Sandbrook et al., 2021).

3.2.7 Using a small dataset and the risks of overfitting

One risk of training with small datasets is the potential for overfitting to occur when there is a decrease in accuracy between the training and validation sets (Choe et al., 2020; Schneider et al., 2020; Shorten & Khoshgoftaar, 2019), or where there is an overly complex model for a limited dataset. Overfitting can result in the model only recognizing some classes when they appear in combination with specific backgrounds seen during training (Schneider et al., 2020), or the model performs well on the training data but not the validation data (Choe et al., 2020; Shorten & Khoshgoftaar, 2019). Consequently, the resultant model cannot generalise well and may exhibit poor accuracy on unseen test data.

There are several mechanisms available to check and reduce the likelihood of overfitting. These are a) visually checking for overfitting, b) using early stopping, and c) using data augmentation.

a) Checking for overfitting

The training and validation accuracy at each epoch can be plotted to check if overfitting has occurred, visually. For a 'good' model, the validation error must continue to concurrently decrease with the training error (Shorten & Khoshgoftaar, 2019).

b) Early stopping

Early stopping can be used, which stops training before the model has overfitted to the training dataset (Schneider et al., 2020), based on a user-defined trigger whereby no changes in a loss metric (usually the validation loss) are observed over a given number of epochs, also known as the patience. We used 100 epochs as the maximum number of epochs each model could run. We used early stopping patience of three whereby if the validation loss remained unchanged after three epochs, training was terminated. Ferreira et al. (2020) also used early stopping patience of three epochs to train on 31 individual classes of three species of birds.

There is a trade-off between the number of epochs and early stopping. On the one hand, some models can take hundreds of epochs to converge, and in theory, the more epochs a model runs for, the better the computer learns the features of all the photos in a dataset. On the other hand, allowing a model to run for several epochs without early stopping can lead to overfitting and be computationally expensive.

c) Use data augmentation

Data augmentation is a technique used to boost limited datasets to possess the characteristics of big data (Shorten & Khoshgoftaar, 2019; Tabak et al., 2018). It involves using either basic image manipulations or augmentation algorithms to produce modified copies of the ground-truth data (Ferreira et al., 2020; Niemi & Tanttu, 2018). Although a respectable amount of data was collected for this specific problem, the TOT_SP_37 dataset is relatively small compared to the size of camera trap datasets, which can contain millions of photos (Beery et al., 2018). It is widely accepted that more extensive datasets result in superior model performance in deep learning (Schneider, 2019; Schneider et al., 2020); therefore, data augmentation is a valuable tool in this case.

3.2.8 Step 7: Using offline data augmentation for class balancing and Step 8. Using online augmentation during training

Data augmentation was used in two ways, to perform oversampling before training (step 7) and to further increase the size of the TOT_SP_37 dataset during training (step 8).

Class imbalance is where there is an unequal class size in a dataset. Data can be oversampled or duplicated to ensure an equal class size in a dataset. Schneider et al. (2020) found that classes with fewer data had a lower recall (proportion of positives identified correctly), than classes with more data. In their review on methods to address the phenomenon of class imbalance, Buda et al. (2018)

found that when training on various deep learning image datasets, such as ImageNet, oversampling was the best method (Buda et al., 2018). Using the imgaug package in Python (Jung et al., 2020) one round of augmentation was performed on each class to avoid multiple copies of the same image. This process is also known as offline data augmentation (Shorten & Khoshgoftaar, 2019). Per image, there was a random choice of seven different tools (flipping, randomly cropping, and blurring images, adding noise to images, slightly changing the colour of images, and affining images, where images could be scaled, zoomed, rotated, or sheared). In terms of generalizability, blurring images could lead to higher resistance to motion blur if present in test images (Shorten & Khoshgoftaar, 2019).

The minimum number of photos considered was 125 per class, though some classes had over 300 photos; post-augmentation. We then randomly removed images to achieve an approximate class size of 250 (ranging from 220 to 260 photos). We used the splitfolders package, a method also employed by Deeb, Roy, and Edoh (2021); after oversampling was performed, the data were randomly split into 70% of the images for training and 15% respectively for the validation and test datasets. After using oversampling, there were 6,282 training images, 1,328 validation images, and 1,386 test images belonging to 37 classes (species).

TensorFlow takes digital image data as input, encoded as a tensor or array of specified dimensions (height x width x colour channels). It allows the efficient loading and resizing of large amounts of data in a generator. When we wanted to investigate the effect of data augmentation more specifically on model performance, we carried out real-time, online augmentation in TensorFlow, which is less computationally expensive. In the training dataset, we rotated images, flipped them, and shifted their width and height. Only augmenting training data allows the augmentation effect to be deduced from the validation and training accuracies.

3.2.9 The structure of neural networks

An artificial neuron in a neural network has several inputs, each with an associated weight. For each artificial neuron, the inputs are multiplied by the weights, summed, and then evaluated by a non-linear function, called the activation function (e.g., Sigmoid, Tanh, ReLu, or Sine) (Tabak et al., 2018). A softmax function is used at the final layer to ensure that the sum of the outputs equals one, i.e., the maximum probability outputted per class is 1 (Tabak et al., 2018).

Convolutional layers can be viewed as applications of a filter to an input that results in activation. Repeatedly applying the same filter to an input results in a map of activations called a feature map, which indicates a detected feature's location in an input, such as an image. In images, these features are maps of pixels. Pooling layers build robustness to minor distortions in images (Niemi & Tanttu, 2018). GAP layers transform an M x M x N feature map to a 1 x N feature map where M x M is the size of the image, and N is the number of filters (Kumar et al., 2021). They reduce the feature map's size, thus minimising the computation time (Kumar et al., 2021; Z. Li et al., 2019). Additionally, GAP layers do not require parameter optimisation, which minimises the risk of overfitting.

The standard CNN, the dominant architecture used for image classification, is built with fully connected layers and several blocks consisting of convolutions, an activation function layer, and pooling layers (Buda et al., 2018). Training and evaluating the performance of CNN's requires significant computational power, which is aided with modern graphical processing units (GPU's) (Buda et al., 2018). Generally, the input sources for CNN architectures are an image with RGB (red, blue, green) pixel channels and fixed size (for example, 224 x 224 pixels). The output is the predicted species class, as for many other camera trap image-based species classification models (Schneider et al., 2020).

3.2.10 Hyperparameter fine-tuning

Hyperparameters are parameters whose value is used to control the learning process. When training image recognition networks, these are variables that control the number of images used in every epoch (the batch size), minimize error (the optimiser), and determine how a network is trained (the learning rate) (Kandel & Castelli, 2020). Different configurations of hyperparameters can be tested to see which result in the highest accuracy. Given its connectedness and feed-forward nature (Couret et al., 2020), the DenseNet network was used to investigate the effects of batch size and optimiser selection, with no pre-processing tools such as data augmentation.

First, we tested the effect of varying the batch size. The batch size and the optimiser interact, as each batch is fed through the network, based on the results from an 'objective function' (i.e., the average answer accuracy across a dataset), the optimiser modifies the weight values to improve accuracy (Tabak et al., 2018). Small batch sizes are said to improve models' generalisation capacity (Ferreira et al., 2020). In contrast, large batches can reach better, more general minima (giving the model a greater opportunity to learn as more data is processed at each epoch) than when using a small batch size (Kandel & Castelli, 2020).

Next, we tested the effects of different optimisers based on their default learning rates. These were the adaptive momentum algorithm (Adam), a standard optimiser function, known for its speed, the Stochastic Gradient Descent (SGD), proven to be an effective way of training deep networks (Ioffe & Szegedy, 2015), the AdaGrad optimiser (similar behaviour to SGD) and the RMSProp optimiser. Adam has been used in other applications of machine learning to bird species identification from photos (Choe et al., 2020; Das & Kumar, 2018; Ferreira et al., 2020). Nevertheless, it may not be the most suitable optimiser for our data, as there is no indication that one optimiser may outperform another before being applied to a data problem. Once the best hyperparameters were identified, we trained each network from Table 3.1 on the TOT_SP_37 dataset.

3.2.11 10. Training models which are robust to occlusion

CNN's which are trained on visible objects may fail to recognise objects which are partially occluded or obscured (Beery et al., 2019; Schneider et al., 2020). Hence, to account for the possibility of occlusion, it is crucial to train on data that contains at least partial occlusions.

Hughes and Burghardt (2017) investigated the effect of partial occlusion by waves on individual detection of great white sharks; though they did this by randomly assigning a specified number of ‘lower visibility’, partially occluded images to both the training and validation sets. Ensuring an equal distribution of occluded images across training, validation and test splits is one way to guarantee a model is robust to occlusion. However, there are other, more complex image analysis techniques to achieve this.

We tested the ability of models to deal with occlusion using random erasing (Zhong et al., 2017), which is a form of data augmentation. In random erasing, each image within a batch randomly undergoes two possible operations; either it remains unchanged or an area of the image is masked with randomised pixels (Shorten & Khoshgoftaar, 2019; Zhong et al., 2017). This makes the model more robust to noise and occlusion in images during training. Other techniques for dealing with occlusion were tested, such as image inpainting, though this had a long inference time so was not feasible for this study (see Appendix II Section 1 and Figure A2.4 for further details). It is a potentially interesting area for dealing with occlusion, but is currently not scalable, given the limited resources we had at our disposal (i.e., limited access to supercomputers).

We used random erasing with pixel-level randomisation and the best hyperparameter configuration used by Zhong et al. (2017), relating to the probability of applying random erasing (applied to 50% of the images in each batch) and the area and aspect ratio of the boxes. These hyperparameters caused the greatest decrease in test error rate when training on the CIFAR-10 dataset (a benchmark dataset consisting of 60,000, with 6,000 images per class).

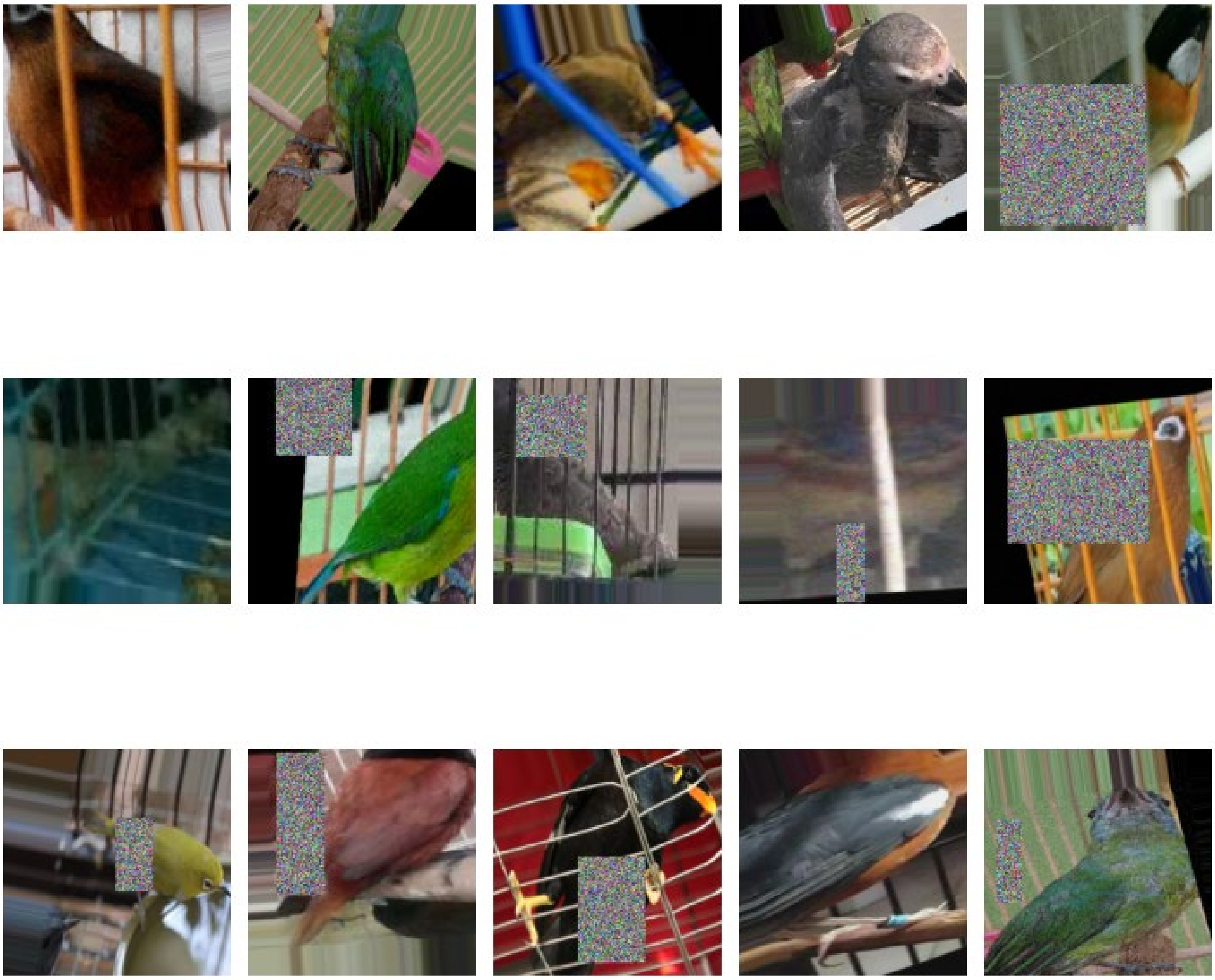


Figure 3.2: The effect of random erasing when applied to 15 images in a batch, with a probability of random erasing being applied of 0.5

3.2.12 Using transfer learning to train image-based multiclass classification models

There are two ways to use transfer learning. Either transfer learning can be transformed from pre-trained weights and initialize the model with pre-trained ImageNet weights or use weights from other benchmark datasets. Alternatively, some of the high-level layers of a feature extractor can be unfrozen and fine-tuned, with the final layer which predicts the classes.

We first tested the performance of networks used in similar image-recognition classification problems (e.g., Choe et al., 2020; Couret et al., 2020; Das & Kumar, 2018), and newer classes of machine learning architectures, such as transformers. The Vision Transformer (recently released by

Google) interprets images as a series of square patches, then feeds them to the model in a sequence of flattened 2D patches. Learnable position embedding's are added to each patch to allow the transformer encoder to learn about the original structure of images (Dosovitskiy et al., 2020).

Out of the previous work described in the Introduction, we did not test the support vector machine, the ResNet-50, or the VGG-19 architectures. The support vector machine network (Niemi & Tanttu, 2018) is complicated to ensemble with other models, which was omitted from our experimentation. Fink et al. (2021) did not re-train the ResNet-50 model on their data, so it was not selected for the task of transfer learning. Lastly, the VGG-19, used by Ferreira et al. (2020), was not used as it is slow to train, the network architecture weights are too large (above 500 MB), and it is difficult to compress for applications.

Table 3.1: The different machine learning architectures introduced in the Introduction tested on the TOT_SP_37 dataset.

Type of architecture	Name of model	Memory Size	Parameters	Examples from the literature
CNN	Inception V3	92 MB	23,851,784	Das and Kumar (2018)
CNN	DensetNet121	33 MB	8,062,504	Couret et al. (2020)
CNN	Densenet201	80 MB	20,242,984	Couret et al. (2020)
CNN	NasNetMobile	20 MB	5,326,716	Choe, Choi and Kim *(2020)*
Transformer	Vision Transformer (VT)	≈ 417 MB (398 MiB)	87,467,239	Dosovitskiy et al. (2020)

3.2.13 Model structure

Before models were trained, we set a random seed for reproducibility. A random seed ensures that results can be reproduced if the model needs to be re-run.

Each architecture from Table 3.1 was initialised with ImageNet weights, minus the top layer, known as a 'headless' model. For all CNN's, the headless model was stacked with a global average pooling layer (GAP), a fully connected Dense layer (size 1024, with ReLu activation function). The last layer of each model was replaced with a Dense layer of size 37, following the number of classes (bird species). The activation of the classifier layer was controlled by the softmax function. This function allocates decimal probabilities to each class, which add up to one.

An example structure was followed for the vision transformer model (available at https://www.kaggle.com/raufmomin/vision-transformer-vit-fine-tuning), whereby the headless model was stacked along with a combination of flattening, batch normalization, and dense layers. Flatten layers flatten the inputs, whereas batch normalization makes the network more stable during training (Kandel & Castelli, 2020). Batch normalization may require the use of much larger than standard learning rates, which speeds up the learning process. We also used the Rectified Adam optimiser, as outlined in the linked example, which converges in less epochs than Adam does (Liu et al., 2020).

The Inception V3 network was fine-tuned to see if fine-tuned models performed better than frozen, headless models. The top two inception blocks were unfrozen, so the first 249 layers were frozen, and the rest were given the opportunity to tune and learn on the data.

To further investigate ways to improve the model accuracy, all architectures were ensembled, by averaging each model's output before compiling and training the model. Another way to do this is to train each model and then average their predictions (Norouzzadeh et al., 2018). It is of value to evaluate the performance of multiple models together since an ensemble of models may further improve classification accuracy (Norouzzadeh et al., 2018; Sagi & Rokach, 2018).

3.2.14 Performance metrics

All models were queried for training, validation and test accuracies. The accuracy equals the proportion of correct model predictions out of the total number of predictions. Both the batch test accuracy and whole test accuracy were calculated. The batch test accuracy calculates the proportions of correct predictions in a randomised batch from the test dataset. The whole test accuracy calculates the accuracy rate across predictions for the whole test set and represents the difference between the true labels and the predicted ones (Schneider et al., 2020). The *precision*, *recall*, and *F1 scores* were also computed on the test data. The *F1 score* (also known as the harmonic mean) is calculated from the precision and recall (Schneider et al., 2020). The F1 score is a more valid measure of model performance than the test accuracy, as it gives a better measure of the incorrectly classified cases.

$$Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{recision} + \text{recall}}$$

3.2.15 Evaluation metrics

A confusion matrix was plotted for the best model to further evaluate the model performance on each species. The confusion matrix shows the true positives, true negatives, false positives, and false negatives between classes (Schneider et al., 2020). This can be used as a diagnostic tool to identify species commonly confused with one another.

If the performance of the single train-test split can be widely retained after repeated testing, this suggests that the model generalizes reasonably within the domain tested, which in this case was primarily wildlife markets (Sakib & Burghardt, 2021). However, it is possible to accidentally train on a subset that does not reflect a real-world situation. The validation or test datasets could consist of easier examples than the training dataset. By training the model on randomly shuffled training, validation, and test datasets, any tendencies towards overfitting can be avoided. The similarity of the folds across the training, validation, and test datasets in the TOT_SP_37 dataset can be observed in Tables A2.3-A2.5).

The best-performing model was evaluated using stratified 5-fold cross-validation. The splitfolders package (Filter, 2020) was used, without a seed, to generate five random data folds, each with a 0.7, 0.15, 0.15 data split. Due to the novelty of this network (released at the end of 2020), 5-fold cross-validation was also performed on the vision transformer model.

3.2.16 Understanding the features used by models to perform identifications

Several criticisms of CNN applications are that they are 'black box' applications, opaque to their human users (Durán & Jongsma, 2021; Xin et al., 2018). Another criticism is that the features of images that drive the learning process are difficult to pinpoint.

Following Miao et al. (2018), the gradient-weighted class-activation-mapping (Grad-CAM) procedure was applied, extracting the most salient pixels in the final convolution layer. Grad-CAM offers a partial solution to this opacity by allowing us to see which region of an image most influences predictions and gradients when applying neural networks to our data (Couret et al., 2020; Miao et al., 2018). By inspecting the results, the indirect reasons for the CNN classifications can be determined. As depicted later in Figure 3.9, red and yellow areas in the activation maps depict the regions that the neural network relies on when calculating the prediction (Couret et al., 2020).

3.2.17 Building a Shiny application

The best model was saved in the SavedModel format in TensorFlow, then used to port the model to a Shiny application using the Shiny package (Chang et al., 2020) in the R statistical programming environment. To demonstrate the portability of applications of this nature, we built a graphical user interface (GUI) for our neural network to be run locally through R, using the Shiny package.

3.2.18 Using the application

In the design of the GUI, previous criticism directed to previous apps was factored in. Criticism has focused on too many steps to arrive at an identification (Joanny, 2020) or using decision tree logic, which can be difficult to follow (Kretser et al., 2015). Thus, we provided simple instructions (see Figure A2.5):

1. Take a picture of a bird
2. Crop the image, so the bird fills out most of the image
3. Upload the image (in .jpg, .jpeg or .png format).
4. Observe the top-5 species the model predicts could be in the image

Users can upload images of birds and view the top-5 predictions for what the species are likely to be and the percentage confidence in each identification. We tested the Shiny app on new photos for different categories of photos (caged, uncaged, and wild) for our best model. The tool can not yet be made available online in an unencrypted form, though we present the code for its design. The tool was designed in R version 4.0.4, using the R Studio.

3.2.19 Further understanding the effects of the occlusion on image classification

As aforementioned, heavily occluded images are typically more difficult to classify than unobstructed (Chandler & Mingolla, 2016; Shorten & Khoshgoftaar, 2019; Zhong et al., 2017). It remains to be seen if occlusion causes a significant drop in model performance. To investigate if this is the case for birds, we built a binary classifier to distinguish between 'caged' and 'uncaged' images in our dataset. 'Caged' photos have cage bars obscuring the bird in the foreground, and 'uncaged photos' do not have bars obstructing the view of the bird in the foreground. The model architecture selected for this binary model was VGG-16, to separate photos into caged and uncaged. Although we did not use any VGG networks for training on the TOT_SP_37 dataset, it is more efficient to use on a binary classification problem, particularly with small datasets. We trained this model for 100 epochs, with a batch size of 16, the Adam optimiser, and patience of 10 epochs. Using the pandas package (McKinney, 2020), the predictions from our model were rounded to either 0 (caged) or 1 (uncaged), then exported as a .csv file, which was then used to sort the photos into caged and uncaged datasets. In total, 1,871 images were detected as uncaged of the original images in the TOT_SP_37 dataset (31.38%, 5,963 images).

Once the original dataset was split into caged and uncaged photos, since some species already had a high degree of occlusion, there were not enough uncaged photos that could be used for the artificial caging process. Thus, we only had enough uncaged data for 26 species, hereafter referred to as the "UNCA_26" dataset. Oversampling was performed using the splitfolders package once (performing duplications only once to achieve a standard number of 127 pictures in each training class). Oversampling resulted in 2,666 images for training, 570 images for validation, and 594 images for testing.

All photos in the UNCA_26 dataset were then overlayed with transparent masks of cage bars, which were manually segmented using the GNU Image Manipulation Program, GIMP, version 2.10.22 (GIMP, 2020). Sixty masks were used, 56 of which were generated from real images, whereas four transparent png's were downloaded from Google to introduce further variation, using the search terms 'transparent bars png'. A sample of the manually segmented masks can be seen in Figure 2.2, and Figure 2.3 for the png's downloaded from Google.

We used the splitfolders package (Filter, 2020) to divide both the uncaged and caged datasets (training: 70%, validation: 15%, and test images: 15%). Using a random seed, we split the caged data along the same division, so the model trained on the same ground-truth images, though one iteration was only with artificial cage masks in the foreground. We split the uncaged and artificially caged dataset again to recombine them to make folds of data with 75% of the photos caged, 50% caged and 25% caged. Mixing the uncaged and artificially caged datasets allows us to investigate further the effects of occlusion in the foreground of images

The use of synthetic examples may improve the generalisation of the model (Beery et al., 2018) to classify unseen data with a range of foreground occlusions (i.e., cage bars in various orientations, colours, width, and quality of the mask itself). Further, when training images with various levels of occlusion are generated, this may reduce the risk of over-fitting and make the model robust to occlusion (Zhong et al., 2017).

3.2.20 Hardware requirements and code release

TensorFlow supports training on Graphics Processing Units (GPU's). GPUs are widely utilised in deep learning applications to reduce the computation time for training and inference tasks (Couret et al., 2020). All code was written in Python, version 3.8, using the Keras, an open-source machine learning library that uses TensorFlow as a backend engine.

Information Services at the University of Kent provided specialist and High-Performance Computing (HPC) systems. A custom Docker image was built for TensorFlow 2.4 (available at https://github.com/sam-watts/tf-docker). This image installs the GPU version of TensorFlow, an NVIDIA base image, and the CUDA dependencies for connecting to GPU's. This image is compatible with the NVIDIA Tesla P100 Data Centre GPU, which are commonly those provided by university HPC's

All the code can be run on a consumer-grade laptop on a Windows Operating System (specifically Windows 10 Pro, 64 bit), using the TensorFlow CPU version. When using the TensorFlow GPU version on a consumer-grade Windows PC (in this case, equipped with an NVIDIA GeForce MX150 graphics card), the code will run faster than the CPU version, though considerably slower than using the HPC system.

The code, methodological pipeline and guidance for using HPC's (written in Python version 3.8 and Markdown, via Google Colaboratory) are made publicly available for other researchers to train their models on wildlife-trade-related data and compare results as a community (https://github.com/Sicily-F/cagedbirdID).

3.3 Results

Here rigorous experimentation for a species classifier for birds in the wildlife trade, specifically in Indonesia and China, is presented. The classifier is based on models trained on largely whole-body images of birds, with most photos consisting of a single bird per image.

3.3.1 MegaDetector performance

The MegaDetector performed well on our data. Each species class was processed separately, and we recorded a range of 66% to 98% effectiveness across classes.

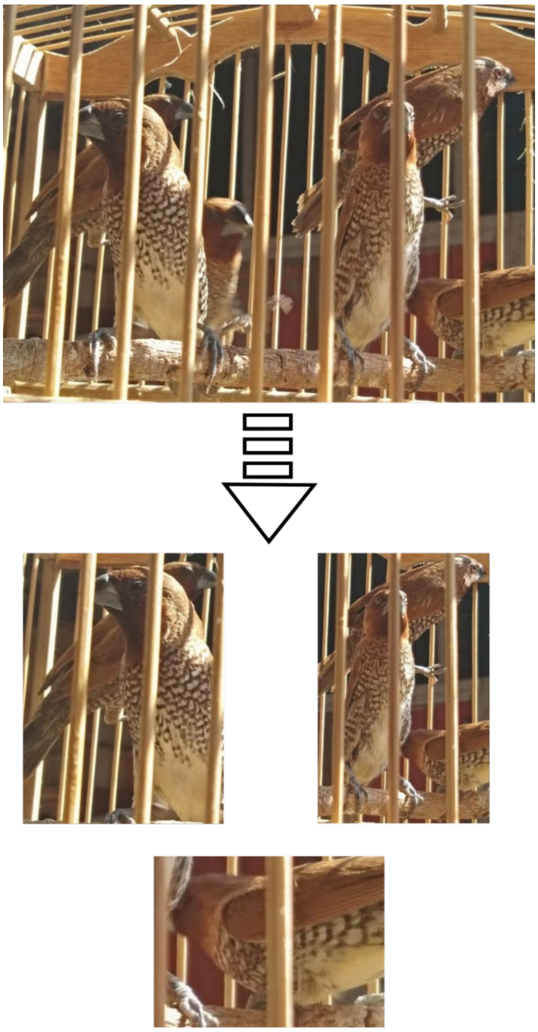


Figure 3.3: The effect of the MegaDetector on cropping an image of multiple caged birds of the species Scaly-breasted Munia, *Lonchura punctulata*.

3.3.2 Hyperparameter fine-tuning

Given our limited resources, we could not test a batch size of 128, as the memory requirements were too large. The batch size of 16 and 32 showed similarly optimal results, though, with a batch size of 32, this required more epochs before stopping. Thus, in a computationally constrained approach, a batch size of 16 was selected for the rest of the experiments.

Table 3.2: The model performance of a DenseNet-121 architecture with different batch sizes, using Adam as an optimiser with a learning rate of 0.00001. All values are to 3 decimal places.

Batch size	Training accuracy	Validation accuracy	Batch test accuracy	Number of epochs the model ran for
8	0.999	0.911	0.875	20
16	**0.999**	**0.918**	**0.938**	**22**
32	1.000	0.896	0.938	26
64	0.99	0.890	0.891	38

All the optimisers returned a good batch test accuracy, however, out of the optimisers we tested, the SGD optimiser resulted in the best test accuracy in the shortest number of epochs.

Table 3.3: The model performance of a DenseNet-121 architecture with different optimisers, with TensorFlow default learning rates, using a consistent batch size of 16 and a patience of 3 epochs. All values are to 3 decimal places. lr is the learning rate.

Optimiser	Hyperparameters and requirements	Training accuracy	Validation accuracy	Batch test accuracy	Number of epochs the model ran for
AdaGrad	lr =0.001, initial_accumulator_value=0.1, epsilon=1e-07	0.991	0.913	0.875	30
RMSProp	lr=0.00001	1.000	0.944	0.938	19
Adam	lr=0.00001	0.999	0.911	0.975	24
SGD	**lr=0.01, momentum=0.0**	**1.000**	**0.922**	**0.875**	**10**

3.3.3 Training different architectures

One goal of this chapter was to train different machine learning architectures to distinguish each of the classes in the TOT_SP_37 dataset. In Table 3.4, we can also see the effects of the different pre-processing methods, which help to make a model robust to occlusion.

We consistently achieved whole test accuracies above 0.75 across all networks, with a batch size of 16, in all networks, except for NasNetMobile, and whole test accuracies of 0.85. Although Inception V3 has a deeper network structure than some of the other models, it performs slightly worse on the task of bird identification (Table 3.4). However, the NasNetMobile model performed the worst of all the networks (Table 3.4).

Table 3.4: The model performance of four neural networks, one transformer, and one ensemble model, with a batch size of 16, using the SGD optimiser and early stopping patience of 3 epochs. All values are to 3 decimal places, with the exception of total test accuracy, which is 2.

Pre-processing methods	Model	Training accuracy	Validation accuracy	Batch test accuracy	Total test accuracy	Precision	Recall	F1 score	Number of epochs the model ran for
None	D121	1.000	0.926	0.938	0.93	0.931	0.928	0.928	16
	D201	0.999	0.950	0.812	0.93	0.933	0.931	0.931	12
	NasNetMobile	0.982	0.666	0.500	0.68	0.690	0.680	0.680	29
	Inception V3	1.000	0.797	0.750	0.81	0.814	0.808	0.808	11
	VT	0.975	0.915	1.000	0.91	0.915	0.912	0.912	16
	Ensemble								
Data augmentation only	D121	0.987	0.932	1.000	0.93	0.930	0.927	0.927	14
	D201	0.994	0.952	0.875	0.93	0.932	0.930	0.930	16
	NasNetMobile	0.762	0.677	0.562	0.69	0.701	0.696	0.690	28
	Inception V3	0.927	0.849	0.938	0.85	0.862	0.855	0.854	11
	VT	0.928	0.850	1.000	0.89	0.902	0.891	0.891	11
	Ensemble	0.936	0.882	0.875	0.87	0.882	0.868	0.865	77

Table 3.4: Continued.

Pre-processing methods	**Model**	**Training accuracy**	**Validation accuracy**	**Batch test accuracy**	**Total test accuracy**	**Precision**	**Recall**	**F1 score**	**Number of epochs the model ran for**
Random erasing only	D121	0.984	0.913	0.875	0.90	0.909	0.903	0.903	10
	D201	0.994	0.9345	1.00	0.93	0.935	0.931	0.932	13
	NasNetMobile	0.739	0.650	0.312	0.67	0.695	0.670	0.672	19
	Inception V3	0.933	0.831	0.812	0.82	0.833	0.821	0.821	10
	VT	0.965	0.913	0.812	0.91	0.916	0.914	0.913	14
	Ensemble	0.985	0.901	0.750	0.90	0.904	0.898	0.898	84
Data augmentation and random erasing	D121	0.976	0.925	0.969	0.93	0.931	0.928	0.928	18
	D201	**0.965**	**0.941**	**1.00**	**0.93**	**0.941**	**0.934**	**0.935**	**11**
	NasNetMobile	0.673	0.687	0.500	0.69	0.705	0.691	0.687	45
	Inception V3	0.901	0.867	0.812	0.87	0.878	0.871	0.869	19
	VT	0.930	0.921	1.000	0.92	0.924	0.919	0.919	17
	Ensemble	0.9328	0.901	0.938	0.89	0.899	0.892	0.899	63

For all models, applying data augmentation and random erasing in isolation did not always increase model performance in terms of training, validation, and testing accuracy (Table 3.4). However, with data augmentation and random erasing, models generally trained for more epochs before early stopping occurred had higher precision, recall, and F1 scores.

When used in isolation, data augmentation outperformed random erasing (Table 3.4). Combining data augmentation and random erasing for most models did not result in higher precision, recall, and F1 scores. Models using only data augmentation had higher evaluation metrics for test data than random erasing, except for the DenseNet-201 model, which stayed the same, and the Vision Transformer, which has better performance using random erasing (Table 3.4). Out of every model and data pre-processing combination (Table 3.4), the best performing model was the DenseNet-201 architecture, using a combination of data augmentation and random erasing, with the SGD optimiser and a batch size of 16.

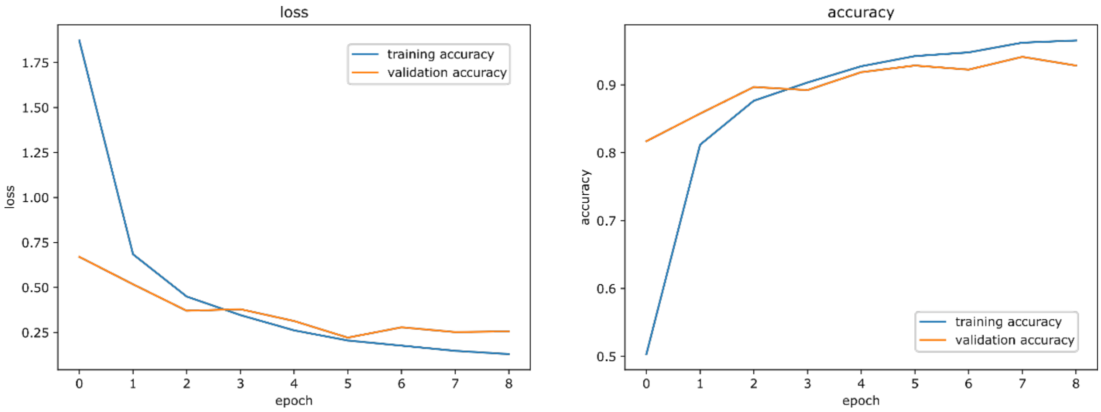


Figure 3.4: A visualisation of the training process for the best model with respect to accuracy and loss. The validation loss was the metric that controlled early stopping, with a patience of 3 epochs.

A concurrent increase in training and validation accuracy can be observed, indicating that overfitting has not occurred (Figure 3.4). Further, there is not a large difference between the final training and validation accuracies.

3.3.4 Using stricter evaluation metrics to assess model performance

The training and validation accuracy appears stable across each fold for the best-performing model (Table 3.5). However, the variation in test accuracy and the number of epochs taken for training suggests that the compilation of each fold and the distribution of photos across training, validation, and test splits, can influence a model's performance. Folds 1 and 3 had a higher total test accuracy (Table 3.5) than our best model in earlier experimentation (Table 3.4). However, the other folds all performed similarly to the best model, trained on one fold.

Table 3.5: The performance of the best model, the DenseNet-201 on, different folds, with data augmentation, random erasing, and a batch size of 16, using the SGD optimiser and early stopping patience of 3 epochs. All values are to 3 decimal places, with the exception of total test accuracy, which is 2.

Fold Number	Training accuracy	Validation accuracy	Batch test accuracy	Total test accuracy	Precision	Recall	F1 score	Number of epochs the model ran for
Fold 1	**0.995**	**0.950**	**1.00**	**0.96**	**0.961**	**0.960**	**0.960**	**20**
Fold 2	0.984	0.941	0.875	0.94	0.938	0.936	0.935	9
Fold 3	0.993	0.943	0.875	0.95	0.946	0.945	0.945	13
Fold 4	0.991	0.933	0.938	0.93	0.935	0.932	0.932	10
Fold 5	0.992	0.950	1.00	0.94	0.947	0.944	0.944	11
Average	0.991	0.943	0.938	0.944	0.945	0.943	0.943	12.6

Table 3.6: The performance of the Vision Transformer on different folds, with data augmentation, random erasing, and a batch size of 16, using the Rectified Adam optimiser and early stopping patience of 3 epochs. All values are to 3 decimal places, with the exception of total test accuracy, which is 2.

Fold Number	Training accuracy	Validation accuracy	Batch test accuracy	Whole test accuracy	Precision	Recall	F1 score	Number of epochs the model ran for
Fold 1	**0.917**	**0.920**	**1.000**	**0.94**	**0.942**	**0.939**	**0.939**	**15**
Fold 2	0.929	0.934	0.875	0.92	0.929	0.924	0.925	16
Fold 3	0.924	0.920	0.938	0.92	0.924	0.920	0.920	19
Fold 4	0.941	0.919	0.938	0.91	0.915	0.911	0.911	14
Fold 5	0.951	0.913	0.812	0.93	0.933	0.931	0.931	19
Average	0.932	0.921	0.913	0.924	0.929	0.925	0.925	16.6

Like our other use of cross-validation (Table 3.6), the training and validation accuracy appears stable across each fold when training using the vision transformer architecture. The average accuracy of 0.924 across folds is approximately the same as the whole test accuracy for the best model on the original dataset.

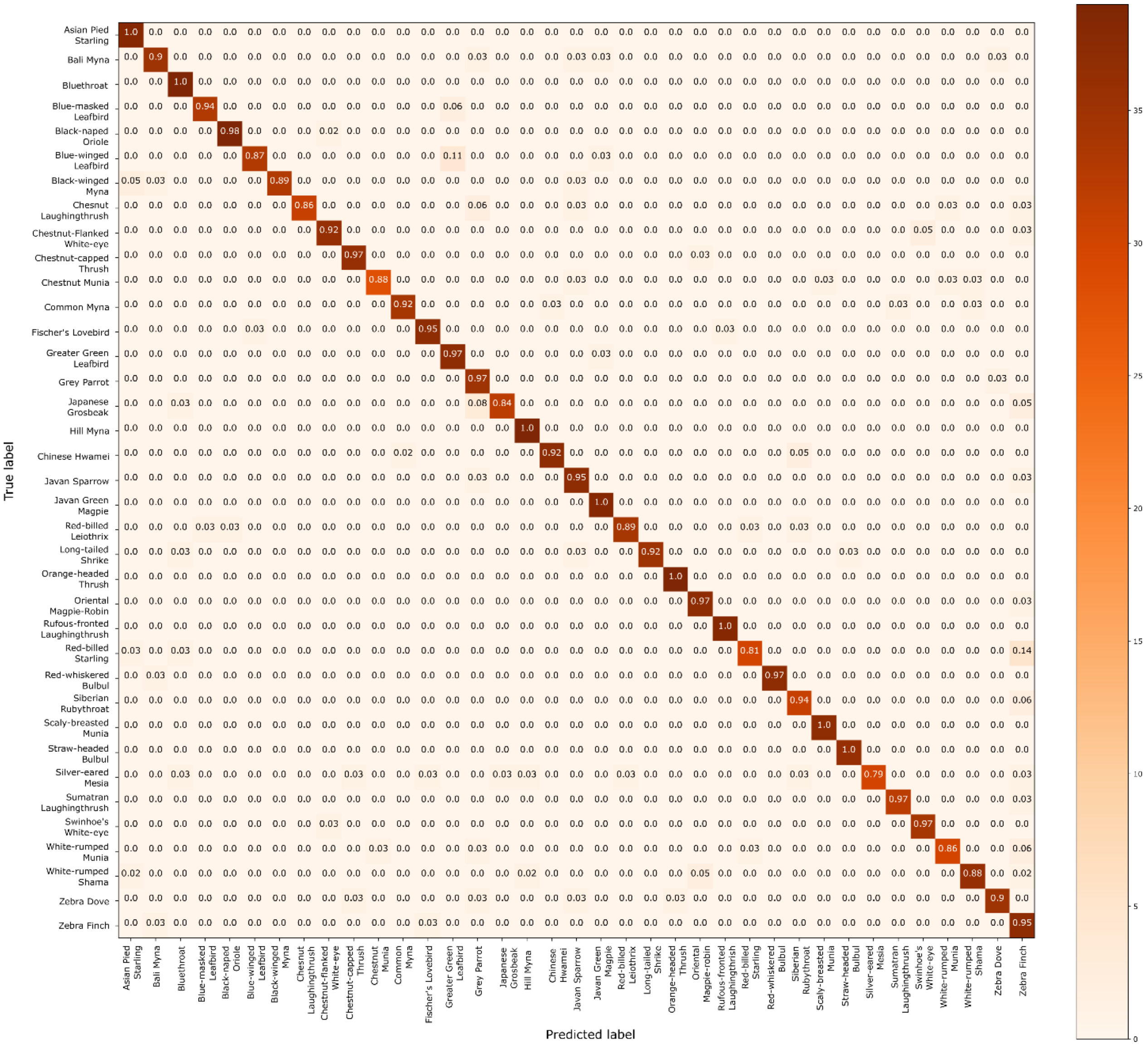


Figure 3.5: The normalised confusion matrix for the best model, the DenseNet-201, on different folds, with data augmentation, random erasing, and a batch size of 16, using the SGD optimiser and early stopping patience of 3 epochs. True labels are on the y-axis, whereas predicted labels are on the x-axis. The higher colours indicate higher recall; the units for the colour scale are the support number of photos for each class in the test dataset.

The higher the confusion matrix's diagonal values (the recall), the better, indicating correct predictions. The highest recall (1.00) was recorded for several species. These were the Asian Pied Starling, *Gracupica contra,* the Bluethroat, *Luscinia svecica,* the Hill Myna, *Gracula religiosa,* the Javan Green Magpie, *Cissa thalassina,* the Rufous-fronted Laughingthrush, *Garrulax rufifrons*, the Scaly-breasted Munia, *Lonchura punctulata* and finally, the Straw-headed Bulbul, *Pycnonotus zeylanicus.* Misclassifications by our best-performing model were rare. Misclassifications were highest for the Silver-eared Mesia, *Leiothrix argentarius* (recall = 0.79). The complete classification report for each species (precision, recall, and F1 score) can be found in Table A2.2.

3.3.5 Applying the Grad-CAM algorithm

The features used by our model for identification can be viewed for five species below, which were chosen due to their varying pose.

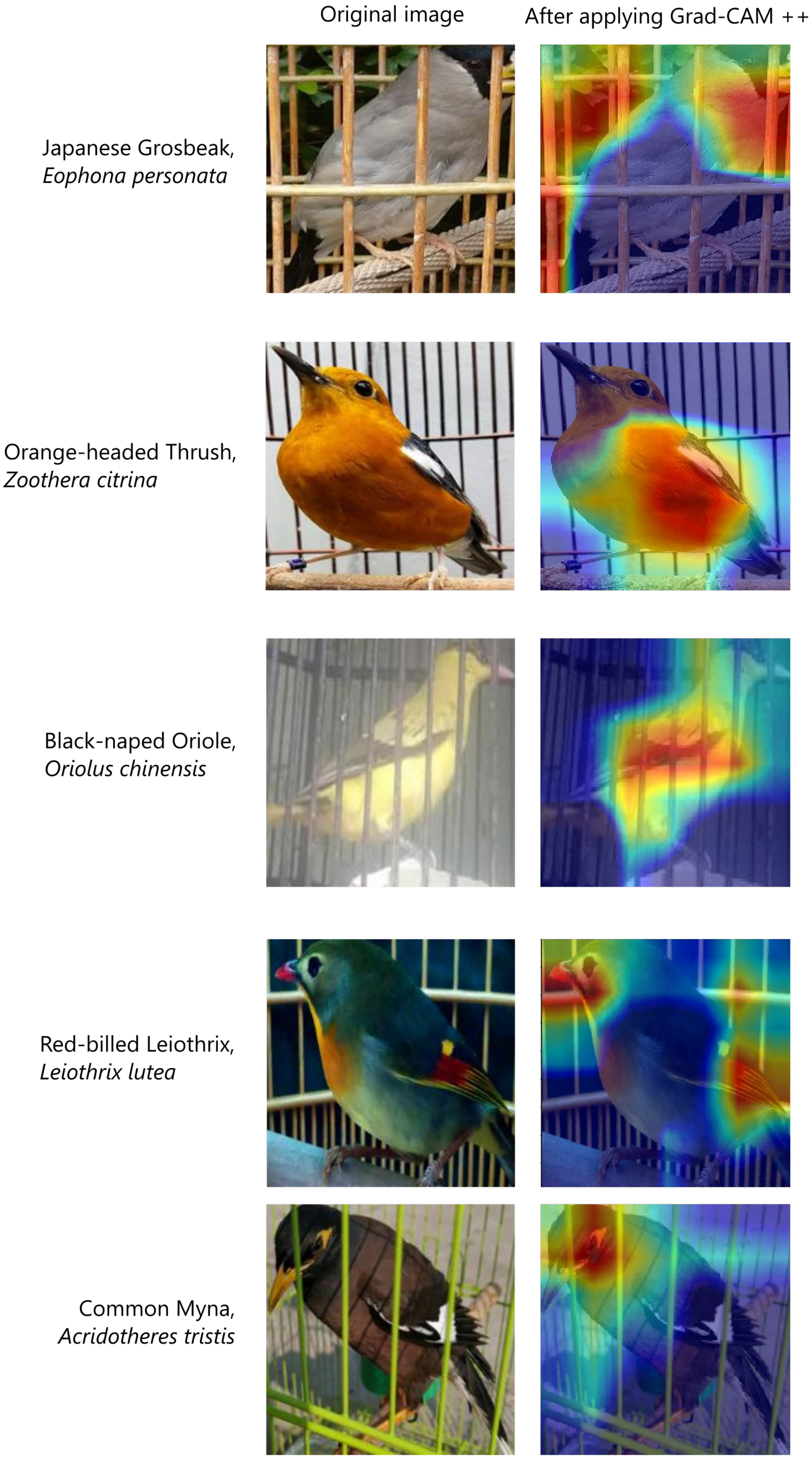


Figure 3.6: The Grad-CAM image shows the heat map, highlighting the region of ‘most interest’ to the CNN.

The DenseNet-201 network places its attention more heavily on the head and wings regions of the birds when calculating the prediction (Figure 3.6). In some cases, such as the Red-billed Leiothrix, both the head and wings were important. For the Japanese Grosbeak, part of the cage was highlighted as relevant to the prediction (Figure 3.6). This highlighting is likely erroneous and could be corrected with more training and or data.

3.3.6 The Shiny application

This model was ported to a Shiny app. The results of the model classification, using the Shiny app, are presented here for two different scenarios: caged and wild cropped. The uncaged cropped, caged, uncaged, and wild uncropped scenarios can be found in Figures A2.6-8.

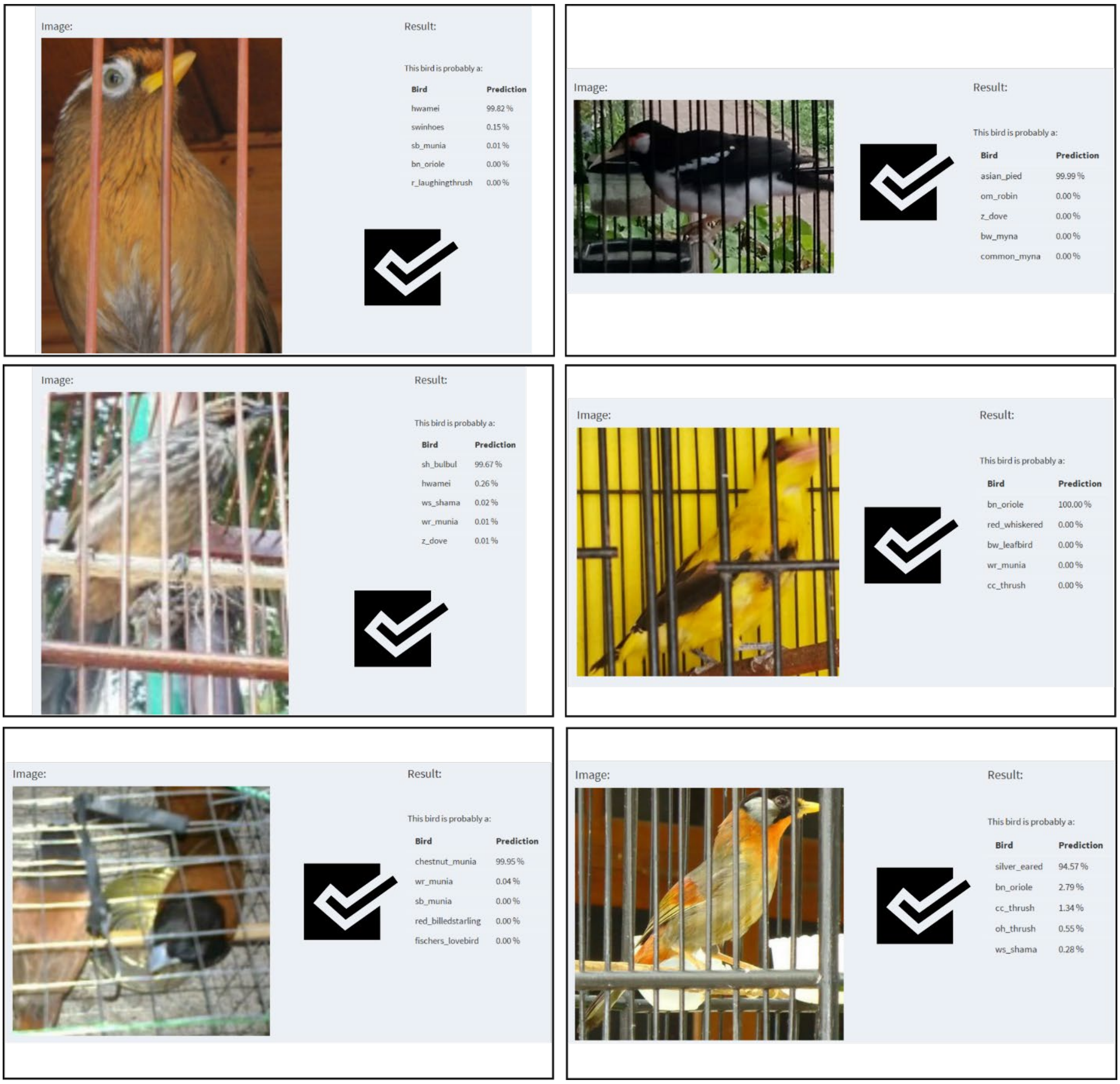


Image:

Result:

This bird is probably a:

Bird	Prediction
hwamei	99.82 %
swinhoes	0.15 %
sb_munia	0.01 %
bn_oriole	0.00 %
r_laughingthrush	0.00 %

Image:

Result:

This bird is probably a:

Bird	Prediction
asian_pied	99.99 %
om_robin	0.00 %
z_dove	0.00 %
bw_myna	0.00 %
common_myna	0.00 %

Image:

Result:

This bird is probably a:

Bird	Prediction
sh_bulbul	99.67 %
hwamei	0.26 %
ws_shama	0.02 %
wr_munia	0.01 %
z_dove	0.01 %

Image:

Result:

This bird is probably a:

Bird	Prediction
bn_oriole	100.00 %
red_whiskered	0.00 %
bw_leafbird	0.00 %
wr_munia	0.00 %
cc_thrush	0.00 %

Image:

Result:

This bird is probably a:

Bird	Prediction
chestnut_munia	99.95 %
wr_munia	0.04 %
sb_munia	0.00 %
red_billedstarling	0.00 %
fischers_lovebird	0.00 %

Image:

Result:

This bird is probably a:

Bird	Prediction
silver_eared	94.57 %
bn_oriole	2.79 %
cc_thrush	1.34 %
oh_thrush	0.55 %
ws_shama	0.28 %

Figure 3.7: The performance of the Shiny application applied to a sample of caged, cropped images, along with the top-5 accuracy.

The Shiny app performs well on images that fit the main class of images fed to the model, namely caged, cropped images. All the classifications in Figure 3.7 were correct, with over 94% confidence in the species prediction for each one.

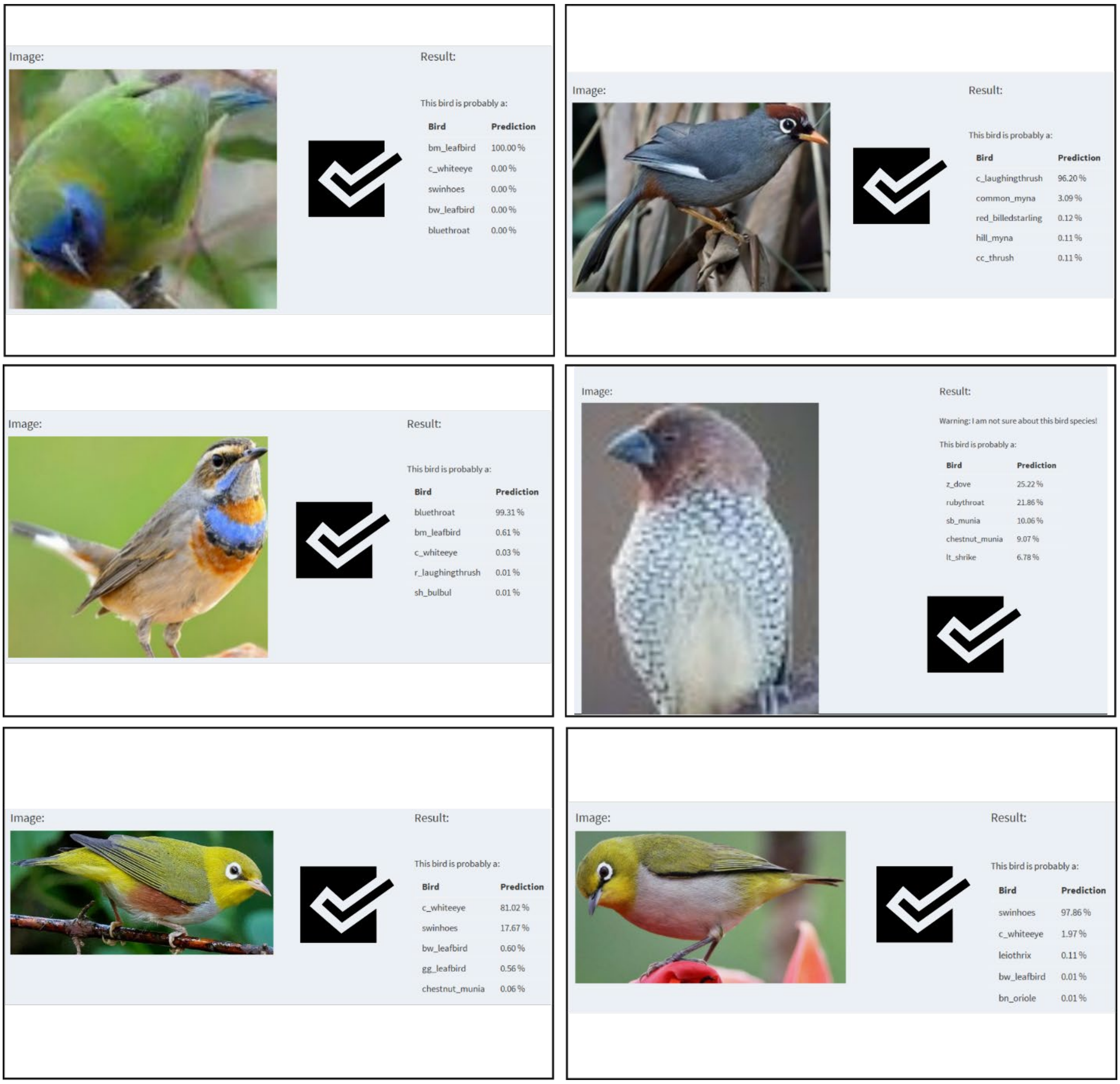


Image:

✓

Result:

This bird is probably a:

Bird	Prediction
bm_leafbird	100.00 %
c_whiteeye	0.00 %
swinhoes	0.00 %
bw_leafbird	0.00 %
bluethroat	0.00 %

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
c_laughingthrush	96.20 %
common_myna	3.09 %
red_billedstarling	0.12 %
hill_myna	0.11 %
cc_thrush	0.11 %

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
bluethroat	99.31 %
bm_leafbird	0.61 %
c_whiteeye	0.03 %
r_laughingthrush	0.01 %
sh_bulbul	0.01 %

Image:

Result:

Warning: I am not sure about this bird species!

This bird is probably a:

Bird	Prediction
z_dove	25.22 %
rubythroat	21.86 %
sb_munia	10.06 %
chestnut_munia	9.07 %
lt_shrike	6.78 %

✓

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
c_whiteeye	81.02 %
swinhoes	17.67 %
bw_leafbird	0.60 %
gg_leafbird	0.56 %
chestnut_munia	0.06 %

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
swinhoes	97.86 %
c_whiteeye	1.97 %
leiothrix	0.11 %
bw_leafbird	0.01 %
bn_oriole	0.01 %

Figure 3.8: The performance of the Shiny model applied to a sample of wild, cropped images, along with the top-5 accuracy.

The model also returns 100% of classifications correctly for wild backgrounds, despite the model being trained on minimal 'wild' backgrounds.

3.3.7 Testing the effects of occlusion on partitioned datasets

We tested how the presence of occlusion affects the accuracy of the model in deciding species identification. In terms of 100% caged versus 100% uncaged, the performance metrics were higher for the 100% uncaged set. Interestingly, the results were better for the 75% caged and 25% uncaged set than the 75% uncaged and 25% caged dataset. The best performance was observed for the 50-50 dataset (Table 3.7), likely due to the equal distribution of the two main classes.

Table 3.7: The performance of the best model, the DenseNet-201, on datasets with differing degrees of occlusion, using data augmentation, random erasing, and a batch size of 16, with the SGD optimiser and early stopping patience of 3 epochs. All values are to 3 decimal places, with the exception of total test accuracy, which is 2.

Degree of occlusion	Training accuracy	Validation accuracy	Batch test accuracy	Whole test accuracy	Precision	Recall	F1 score	Number of epochs
100% caged	0.968	0.916	1.000	0.90	0.910	0.901	0.901	19
100% uncaged	0.956	0.839	0.875	0.81	0.854	0.809	0.812	8
50% uncaged and 50% caged	0.989	0.977	1.000	0.97	0.976	0.975	0.975	14
75% uncaged and 25% caged	0.955	0.921	1.000	0.88	0.907	0.885	0.997	9
75% caged and 25% uncaged	0.983	0.798	0.688	0.66	0.785	0.661	0.675	11

3.4 Discussion

A gap in existing technology is that although some apps are aimed at wildlife trade in Southeast Asia, most have not explicitly focused on songbirds. The current research uses transfer learning to fine-tune existing convolutional and patch-based networks to recognise bird species. To the best of our knowledge, this is a novel application of deep learning. These results demonstrate that deep neural networks can be trained with high accuracy using relatively small datasets to identify species of birds in the wildlife trade.

3.4.1 Applying the MegaDetector

The application of a class-agnostic detector such as the MegaDetector to crop photos was highly successful. Despite not being trained on photos of birds in cages, it was able to generalise well to our dataset (66% to 98% effectiveness across classes). In some cases, the MegaDetector 'created' more images by cropping many birds from a single image (i.e., like Figure 3.3). The creation of images is a beneficial side effect, as it can boost the size of ground-truth datasets. This success may be relatively expected, as although the backgrounds in our data were very different from the backdrops of natural habitats, the MegaDetector is accustomed to occlusion from other natural sources such as rocks, foliage, and other animals (Beery et al., 2019).

These results demonstrate the domain adaptation potential of the MegaDetector, especially since we were able to use it as a cropping tool, not just on photos of birds in the wildlife trade. We also used the MegaDetector to isolate birds from screenshots of social media posts and other non-natural landscapes that animals may occur in. In the interest of pursuing fast image processing solutions, we did not fine-tune a pre-trained Mask RCNN model, also trained on the MS Coco dataset, an approach also used by both Das and Kumar (2018) and Ferreira et al. (2020) to extract the regions of interest for photos of unoccluded birds. For a heavily occluded dataset, such as TOT_SP_37, the MegaDetector still performed well in identifying the regions of interest in our photos, without the need for further training.

However, the MegaDetector performance was less good for some photos of crowded cages, noticeably for the White-rumped Munia, the Chestnut Munia, and the Scaly-breasted Munia, which still suggests a limited application of the frozen MegaDetector model. Either no birds were detected from these photos, or only a few individuals could be cropped from cages containing upwards of 20 birds. With further training, the MegaDetector could extract more animals in high-density environments.

3.4.2 Varying other hyperparameters

There was an apparent effect on the accuracy when the batch rate (Table 3.2) and optimiser (Table 3.3) were changed. However, the reasons why some optimisers work better for specific data or problems are relatively unclear. This lack of clarity indicates the importance of experimentation with optimisers to configure the best model. Testing optimisers, to some degree, becomes a trial-and-error process, and more work can be conducted to vary other hyper-parameters such as the learning rate.

3.4.3 Discussion of transfer learning and experimenting with different architectures

The best model (DenseNet-201, SDG optimiser, batch size of 16) had the highest whole test accuracy, precision, recall, and F1 score (Table 3.4). Like Willi et al. (2019) these results show the increasing benefit of transfer learning for smaller training datasets. Despite DenseNet-201 having the best performance, many other architectures performed well, apart from the NasNetMobile model. There are several explanations for why model performance using other networks may have been so high. The high-test accuracy from our best model is likely because of mechanisms to increase the size of the TOT_SP_37 dataset (data augmentation) and to make the model robust to occlusion (random erasing), which combine to render a strong model performance.

These results also converge with other studies with small training datasets, which used different CNN networks. For example, Hernández-Serna and Jiménez-Segura (2014) achieved 92.87 % training accuracy with only 1,800 training images for 32 species of fish. Hence, it is possible to harness the power of transfer learning for classifiers with small datasets. In comparison, TOT_SP_37 contained nine times more ground-truth data than Hernández-Serna and Jiménez-Segura (2014), with a training accuracy of 96.8%.

Schneider et al. (2020) reported a training accuracy of 95.6% and an F1 score of 0.794 for a DenseNet201 model trained on 55 classes of species from known locations in Canada. Although our model had fewer classes than Schneider's, our best model had a higher training accuracy and F1 score (training accuracy of 0.965 and a recall of 0.934). Couret et al. (2020) also achieved a similar accuracy, to ours, for identifying mosquito species using DenseNet201. However, our model contained more classes and variation in photo quality than their study. One drawback of using the DenseNet-201 model is that it requires a lot of memory (Table 3.1), so it might not be the most practical network to use where memory constraints are a bottleneck for implementing of machine learning models.

The vision transformer also performed well on the TOT_SP_37 dataset (Table 3.2), highlighting a transformer's utility rather than a neural network approach. The results from the cross-validation of this network using 5-folds show a consistency in the test accuracy (Table 3.6). Although this was not done here, in the network's publication, the mean and standard deviation of the training accuracy were averaged over three fine-tuning runs (Dosovitskiy et al., 2020) on different image classification benchmark datasets. The batch test accuracy is more variable (Table 3.6). Given its novelty, it may be that single test results from the vision transformer are relatively unstable and that cross-validation or averaging over three runs are necessary to utilise the model in an application reliably.

As previously mentioned, with early stopping, there is a trade-off between overfitting to a small dataset and the specific model behaviour. If a model runs for too many epochs, it may not generalise well to new data (using small datasets). In contrast, on the other hand, a deeper model (i.e., Inception-V3) or a model which performs well on a small number of classes (i.e., NasNetMobile) might need more time to train, which may partly explain the poor performance of these models.

Surprisingly, despite promising results achieved by Das and Kumar (2018), the Inception V3 architecture did not perform as well on our data (Table 3.4). This may be because of the previously described trade-off between the model runtime and the risk of overfitting. If we had used less cautious patience (more than three epochs), better model performance may have been observed if the model had run for more epochs. However, Buschbacher et al. (2020), who used CNN's to identify wild bee species, also reported lower accuracy with typically 'deep' neural networks such as Inception V3. The Inception V3 architecture can be redeployed when more data can be obtained for this data problem.

The NasNetMobile performed with worse resolve than the other networks. Although Choe et al. (2020) achieved their highest model performance with NASNetMobile, we did not observe such a high performance (Table 3.4). This may partly be explained by the fact that the NasNetMobile model has a low number of parameters that need more time to learn the features of a new set of images. Alternatively, we had more classes of birds than Choe et al. (2020), so the network may be better suited to modest bird classification problems.

That said, we decided to test its capabilities in the ensemble model further. The ensemble model, with data augmentation and random erasing, trained for 63 epochs (Table 3.4), with a much higher final model performance than any NasNetMobile model. Hypothetically, an ensemble would provide a modest improvement in classification capabilities (Schneider et al., 2020). However, the ensemble model did not outperform the individual DenseNet model (Table 3.4). The ensemble model had notably poor test accuracy when random erasing was applied. This may be a by-product of fitting an overly complex ensemble model to a small data problem.

Although we have shown that pre-trained models can perform well on small datasets, TOT_SP_37 also contained relatively few classes compared to the base models, which are trained from scratch on 1,000 classes (Couret et al., 2020). Hence, accuracy may only be temporarily high since the model only has a small number of classes to predict from a comparatively small test dataset. The test dataset was also only 38 photos per class, so the model has few predictions to make. In some cases, the whole model accuracy was higher than the batch test accuracy (Table 3.4). This is relatively unexpected, though this may be due to the random generation of a particularly poor batch. Even for models which had high whole test accuracy (Table 3.4), if we deployed the model in a market, it might not perform as well in a real-world setting because these images are not as clear as the test dataset in the TOT_SP_37 dataset. Alternatively, the test dataset may have less complexity than the training dataset. Hence, for smaller datasets such TOT_SP_37, it may be more appropriate to use the F1 score as the definitive metric for test performance, as well as the average of our 5-fold cross-validation scores. Our model performance can be viewed as provisionally high, and this performance may not persist if we had more classes.

Another explanation for extremely high test accuracy may be because the TOT_SP_37 dataset was biased, despite having photos from a wide variety of backgrounds, contributors, and cameras. Bias may have been inadvertently introduced in the data collection process since the bird had to be clear enough in the photo for a species identification to take place in the first instance. If humans can still identify the birds, they can be theorized as 'easy to identify and in most cases, the head and majority of the bird’s body faces the camera. Further, photos such as those we collected on open-source engines or public social media groups are seldom 'outtake' photos. In addition, it is near impossible to rule out the slight possibility of errors in the manual annotation and labelling of images, which could result in artificially superior model performance.

That said, from the confusion matrix (Figure 3.5), it is relatively unlikely for species to be confused. Although the precision, recall, and F1 score were not perfect for each species (Table A2.2), this information can be used in market surveys to be aware of species easily confused with others. For example, in the case of the Silver-eared Mesia Finch, this can inform targeted data collection for the Zebra Finch, which had a low number of original ground truth images (see Table A2.1). The matrix can also be used to design 'look-alike' species pairings to aid in the future training of customs officers and other stakeholders (Alfino & Roberts, 2019). For example, the Blue-masked Leafbird, *Chloropsis venusta,* was most often confused with the Greater Green Leafbird, a species in the same genus.

3.4.4 Cross validation of models

The outcome of the 5-fold cross-validation procedure (Table 3.5) shows that the best-performing model can classify 37 species of birds, with an average whole test accuracy of 90%. Similarly, for the vision transformer model, we recorded an average whole test accuracy of 0.944 across folds (Table 3.5). The range of test scores for our best model is small, indicating that even on different splits, the model performs well. Using random folds produces more reliable estimates (Table 3.5, 3.6), so the strong performance of our primary model (Table 3.4) could be based on having a fortuitous split. However, a disadvantage is that running on multiple folds increases training costs as the model must be trained on each fold. Further, even though splits were randomly generated, there was considerable overlap between training folds (Table A2.3), though not in the validation (Table A2.4) and test (Table A2.5) datasets. Thus, the evaluating model performance on folds should focus on downstream metrics such as the test accuracy, precision, recall, and F1 score.

3.4.5 The performance of the Grad-CAM

Our initial application of the Grad-CAM procedure, appears to highlight the most discriminative image areas (Miao et al., 2018). In the cases presented (Figure 3.6), these seem to be either particularly colourful parts of the bird or where there are distinct colour shifts. It would also appear that markings on the wings and head, especially around the eye, are features that the CNN uses to make a classification. One drawback of using the Grad-CAM procedure is that variations in pose could result in the model relying on a different feature for identification. The Grad-CAM may be less descriptive in certain contexts, so further testing will be necessary, particularly in caged settings (Figure 3.6).

3.4.6 Model generalisability

As well as assessing our model performance on the test set of the TOT_SP_37 dataset, new images were collected via Google Search. The collection of new images allows for the assessment of model performance on entirely independent data. We can determine whether our model can only identify birds in caged photos or from new, unlearned backgrounds such as wild settings surrounded by foliage. The model's strong performance on both the TOT_SP_37 test data and the newer test images (Figures 3.7, 3.8 and A2.6, A2.7, and A2.8 showed that the model generalizes well. Hence, it is applicable for real-world problems (Niemi & Tanttu, 2018), such as identifying birds in the wildlife trade. Further in our examples from the Shiny application, the model got 100% of classifications correct on cropped caged (Figure 3.7) and wild images (Figure 3.8). It also got 100% of classifications correct on uncaged, cropped images (Figure A2.6) and 83% on wild, uncropped images (Figure A2.7). However, the model did not perform as well (67% of classifications correct) on caged, uncropped photos (Figure A2.8).

3.4.7 The impact of occlusion in model performance

In our experiments on model performance, models which used data augmentation and random erasing had higher training, validation, and test accuracy than when no pre-processing methods were employed (Table 3.4). We thus advocate using both augmentation methods to improve the robustness of pre-trained CNN's against occlusion, in this case from relatively linear cage bars.

Chandler and Mingolla (2016) contend that object classification algorithms are less effective as occlusion increases and nearly useless in heavily occluded conditions. However, we did not see an associated decrease in classification accuracy for species with a higher percentage of photos with cages in the foreground (Table 3.7). Even in situations where 100% of the photos are artificially occluded, model performance was still exceptionally high (Table 3.7). Although 60 different foreground cage masks were superimposed onto photos, the models may overfit to a predictable set of foreground occlusions. Accuracy may decrease if more varied foreground masks are used to train on an artificially caged test dataset.

One aspect that we did not measure was the degree of occlusion in the images themselves. For example, if photos were categorized into a slight, medium, or significant occlusion categories, we may observe poorer performance for significantly occluded photos over slight occlusion. Our analyses are based on the presence or absence of occlusion of any degree in the image's foreground. The bird's identity was apparent enough in the photo for an identification to be made during manual annotation, so our results on occlusion should be interpreted cautiously since, in reality, the degree of occlusion may not be so severe.

Although the DenseNet-201 model is robust to occlusion (Table 3.4), occlusion is not a constant in the wildlife trade. In some cases, it is possible to take photos through the cage bars. On the other hand, even if the view of the bird is not obstructed in every photo, employing methods that can learn ‘around’ repeated, occluded objects such as cage bars will be helpful in scenarios of partial image capture (Gomez-Villa et al., 2016). The likelihood of occlusion in wildlife markets is very high. It is likely not possible to spend time observing one bird, which may draw attention and irritate the trader or stall owner if observers put their hands through the bar to take the photo. Consequently, image-based applications which are designed for species identification in markets must be resilient to occlusion.

Notwithstanding, there are several other practical applications of this method in uncaged environments. For example, in their model containing four species of parrot, Choe, Choi and, Kim (2020) did not train on occluded images. They specifically stated that applications of this nature could be used to prevent the smuggling of endangered species in the customs clearance area, presumably where occlusion can be bypassed. Similarly, when monitoring bird trade online, occlusion may be less present. If users sell birds online, they are more likely to take a 'good' photo where the bird looks appealing to any prospective buyers or fellow enthusiasts.

3.5 Conclusions

Often, 'accuracy' is reported in the machine learning literature, though it is not always specified whether this is the training, validation, or test accuracy. This distinction is crucial as it is hard to gauge which architecture is helpful for specific datasets without knowing an architecture's ability to generalize on unseen data. Ultimately, we recommend that those building applications of a similar nature use the precision, recall, and F1 score to evaluate the model's performance rather than the batch test accuracy. Where possible, cross-validation should be used to average model performance across folds, particularly for small datasets.

We demonstrated that the DenseNet-201 neural network is highly appropriate for a subset of birds in the wildlife trade. Less extensive and faster models include DenseNet-121 and Inception V3 also returned good performance metrics, with a lower computation time and memory consumption. However, caution must be applied when using the Vision Transformer, given its novelty.

More broadly, in terms of wildlife trade, we demonstrate that machine learning solutions can be effectively used to monitor trade, even in highly occluded settings. Machine learning has the potential to revolutionize the way the monitoring of bird markets currently operates. Further, given our methodology, it will be easy to modify our model to incorporate more classes to reflect the true diversity of a real bird market. Such a technological approach has vast potential for more widespread monitoring, leading to tailored enforcement efforts. An automated image recognition tool may eventually help mitigate the impacts of the Asian Songbird Crisis by reducing the sale of illegal species in markets or providing a solid evidence base to promote legislative changes to protect species recorded in markets at unsustainable levels.

Chapter 4 Discussion

As previously highlighted, birds are one of the most diverse and traded groups in the wildlife trade. This diversity poses vast challenges in maintaining real-time inventories across wildlife markets and properly training individuals in species identification. In some cases, CNN's have surpassed human performance in image classification tasks, most notably in the ImageNet Large Scale Visual Recognition Challenges (ILSVRC) (Couret et al., 2020; Wäldchen & Mäder, 2018). Comparisons of human and computer performance on wildlife trade-related image classification problems are relatively scarce. However, in identifying birds in wildlife markets, it is unknown if computers can outperform humans.

3.5.1 Comparing the performance of humans versus computers

This chapter will discuss the results from the data chapters regarding our match-mismatch experiment to build a human baseline of identification error using the MA_MIS_19 dataset (Chapter 2) and our CNN model using the TOT_SP_37 dataset (Chapter 3). Further, we will compare the error rates of humans and computers in identification tasks. Informed by the results of Chapter 3, a machine learning model with random erasing and data augmentation was applied to the MA_MIS_19 dataset.

To compare the error rates of humans and computers for 19 species, for human participants, identification accuracy was assessed with a matching task that required same/different decisions for side-by-side pairings of species of birds. For the computer model, transfer learning was employed by re-training an existing convolutional neural network on new, unique classes of bird species. The datasets viewed by both humans and the computer were approximately the same size; in the randomised stimuli provided to participants in Chapter 2, there were 400 images in 200 pairs. In the test set of the MA_MIS_19 dataset, introduced in Chapter 3, there were 443 images.

Table 4.1: A comparison of the human versus computer model performance. All values are to 3 decimal places, with the exception of whole test accuracy, which is 2.

Metric	Humans	Computer
Mean accuracy across all answers	0.928	0.978 (training) 0.875 (unseen test)
Whole test accuracy		0.86
Precision	0.917	0.868
Recall	0.744	0.857
F-score	0.822	0.856

Overall, the computer model outperformed humans in terms of training accuracy and the F1 score, it performed worse than the best human score (Table 4.1). One human participant achieved the highest accuracy. However, we could not say with confidence that they were a super-recogniser (Figure 2.3). We would still need future tests to identify if non-expert super-recognisers of birds can consistently outperform computers and be detected as statistically significant outliers. One benefit of using computers to perform species identification is that fatigue in the learning process is not a factor. It is likely the case that those who have performed surveys in bird markets are also super-recognisers, who represent a gifted yet minute portion of the general population.

The accuracy was plotted within a phylogenetic tree for both the human baseline and the machine learning model (Figure 4.1). As described in Chapter 2, there are no apparent trends of similar accuracies amongst related species. As aforementioned, there are some clusters of similarity, such as the lowest accuracies recorded for the Chestnut-capped Thrush (*Zoothera interpres*, accuracy of 0.90) and Chestnut-backed Thrush (*Z. dohertyi,* accuracy of 0.89) (Figure 4.1A). However, our human baseline may appear somewhat simplistic, as the accuracy is calculated on the number of correct questions for any combination in which the species appeared. Nonetheless, it is still a valid measure of whether humans are better at recognising distinct or similar bird species.

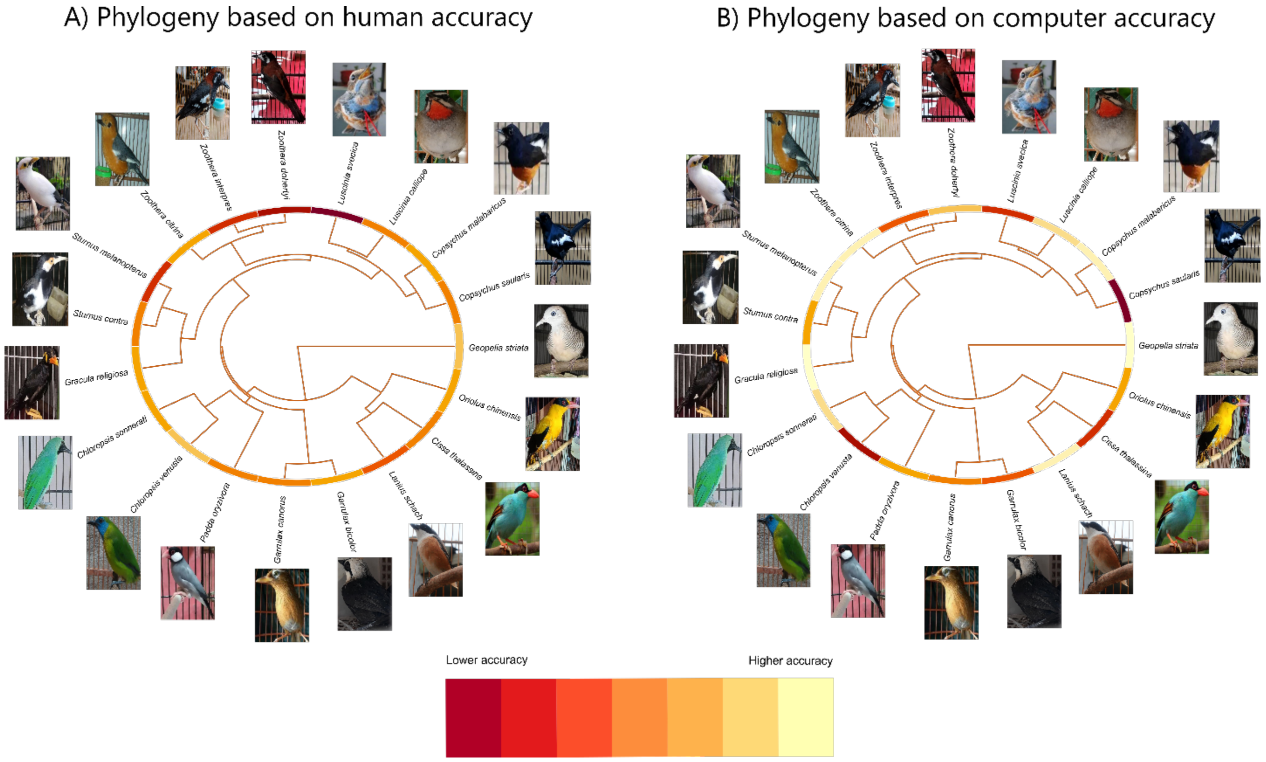


Figure 4.1: A) The phylogeny and the associated accuracy per species for the human classification. B) The phylogeny and the associated accuracy per species can be viewed below for the machine learning model classification.

For the computer-linked phylogeny, accuracies seem more similar amongst related species (Figure 4.1B). However, the phylogenies are not true comparisons of each other as the accuracy was calculated differently. If we had balanced the classes in the Match-Mismatch_19 dataset, the computer might have performed even better. For instance, Schneider et al. (2020) found that classes with fewer data had a lower and more variable recall, although this depended on more significant degrees of magnitude between the classes. We kept the data unbalanced, to intentionally replicate a market situation whereby some birds are in greater abundance than others.

Chapter 2 found that the overall human-based error was 7% in the matching experiment (Table 2.2). Out of the 27 people, four individuals performed significantly poorer than the average score. The computer may be able to outcompete human experts accurately, i.e., if enough data is inputted into a model, it could get 100% of classifications correct in a small subset (as in Table 3.4). Although a computer could achieve 100% in training, in theory, it would be rare for a computer to get all classifications correct on test data. Even the current results from the camera trap literature do not attain 100% on test data (Norouzzadeh et al., 2018; Tabak et al., 2018). Moreover, if 100% was achieved, this may be an artefact of overfitting. However, computers' potential loss of accuracy can be made up for in speed per image processed and 'learned'.

Further, given that the computer accuracy is generated on identifying species rather than telling the difference, we should put more weight on the computer results. Although the computer is dealing with the same decisions, it goes one step further to compare each species via the confusion matrix (Figure 3.5). These results provide a valuable baseline for error rates in both manual and automatic techniques for bird species identification in wildlife markets.

3.5.2 What makes an adequate baseline for human identification and what can they be used for?

A better baseline representation may have been to give our participants a list of the potential species encountered in an experiment, along with images and ask them to define what the species were. This behaviour is roughly analogous to using field guides, a dichotomous key, or a flip chart. Alternatively, we could ask experts to define the species identification and assess their level of consensus. The analysis presented here may not reflect what key features humans use to tell species apart.

In a pre-COVID world, to understand misclassification rates of birds in wildlife markets, a similar matching task, to that detailed in Chapter 2, of pet or bird markets in Southeast Asia could have been trialled with a cohort of law enforcement officers from customs agencies and government environmental protection departments. However, in the absence of this access, psychology methods were utilised, such as match-mismatch experiments. Within the psychology and conservation literature, many studies have used untrained students as proxies for law enforcement personnel (Alenezi et al., 2015; Alfino & Roberts, 2019).

In a scenario where no law enforcement community members have been trained for market surveys or other potential species identification tasks involving birds such as seizures, it would be useful to perform a matching task, such as that presented in Chapter 2. This process could be used a priori to assess which personnel could be relied on to perform accurate species identifications. For trainees who did not score above a certain threshold (for example, less than 10% human error in classification), these trainees can make use of technological assistance such as the machine learning model presented here. These initial tests can identify any super-recognisers; for example, Robertson et al. (2016) suggested that personnel selection optimises performance by taking advantage of any natural super-recognisers in law enforcement agencies.

Although other methods could have been used to create baselines for human identification performance for birds in the wildlife trade, a computerised approach can save valuable time, especially in time-dependent survey situations. Rahimi et al. (2016) posit that computerised systems for taxonomic identification could produce more objective results in less time.

3.5.3 Which features do humans and computers rely on for identification?

In Chapter 3, we displayed the results of the Grad-CAM algorithm (Figure 3.6) to disentangle which areas of images of bird's are used to perform classifications. The application of the Grad-CAM algorithm to training data highlighted exciting clues as to the basis of the DenseNet CNN predictions. Utilising the Grad-CAM results (Figure 3.6), we can reverse engineer what the algorithm learns and highlight critical aspects of a bird's morphology for training humans. Along with the insights from Grad-CAM, the confusion matrix (Figure 3.5), can also act as training material to guide efforts to increase human identification success rates. For example, Miao et al. (2018) trained a CNN to classify 20 African wildlife species, and also applied the Grad-CAM algorithm, and showed that the highlighted pixels in particular images indicated features which in some cases were similar to those used to train humans to identify these species.

In future comparative work, these pixel maps can be compared to newer surveys that identify which features of the bird human participants agree to distinguish it from another image or conversely help match it to the adjacent image. Alfino and Roberts (2019) asked participants to describe the traits they used to identify chameleon species in the *Calumma* genus, and found that most of their participants relied on the nose and head shape to discern species differences. This reliance on one or two features may also be the case for birds. Further work could also involve iris tracking to see where participants gaze for future studies. Iris tracking has been used to examine the effects of pollinator-related labels on consumers' preferences and willingness to pay (Kotsiantis & Kanellopoulos, 2006). The same methods could be used to examine further the role of visual attention in identifying birds.

Gooliaff and Hodges (2018) investigated whether visual features affected agreement amongst experts classifying images of lynx and bobcats. A similar study in the future will help pinpoint the specific features of birds that humans and computers rely on for classification and investigate their potential overlap. Building on Miao et al. (2018) and Couret et al. (2020), we can also produce a visual similarity dendrogram by performing hierarchical clustering of the feature vectors associated with each image in the final fully connected layer in the CNN. However, this was beyond the scope of the current thesis.

3.5.4 Building a case for human-in-the-loop approaches to monitor the wildlife trade

Given our failure to identify any super-recognisers in our cohort and the high test accuracy of the best model (Table 3.4), we contend that a human-in-the-loop approach would improve species identification capabilities amongst experts, increasing accessibility for law enforcement. Human-in-the-loop approaches can be defined as where humans are embedded in machine learning workflows (Xin et al., 2018). The developer 'in-the-loop' uses the end results as cues for modifications to the workflow in general or the architecture. Part of the process of designing artificial intelligence ethically also involves being able to describe what artificial intelligence actually does (Thompson et al., 2021). The term 'human-in-the-loop can also be extended to ensuring that humans using machine learning applications are aware of the aims of using these applications.

There are various challenges if only humans are relied on for identification, even if training has been administered. In face recognition studies, various instruments to increase identification accuracy have been proven to fail. There is no clear link between years of service and task performance, nor the number of training sessions and performance (Robertson et al., 2016). Similarly, taking rests and changing location does not prevent declines in identification accuracy over time, as one would expect (Alenezi et al., 2015). Given all these issues and our comparison of human versus computer performance, a combinatorial approach is likely best.

There are many successful cases of human-in-the-loop identification systems. One famous example is the combination of eGates and trained passport observers at airports, which maximises performance (Alenezi et al., 2015). Further, the FloraGuard project uses a socio-technical AI approach to incorporating human judgement (criminology, conservation science) into iterative cycles of long-tail information extraction (E. Middleton et al., 2020) to analyse online marketplaces for the illegal trade in endangered plants (Lavorgna et al., 2020). We advocate that technological solutions to monitor the wildlife trade should combine the skill of expert participants with automated identification applications, such as the case study presented here. More time and effort must be dedicated to photo labelling and creating accurate species classifiers, which may ease the pressure on those carrying out manual market surveys.

3.5.5 Future benefits of the cagedbirdID application

Machine learning solutions using image processing can improve the confidence of stakeholders in their identifications, as well as be used as a survey aid by less experienced and less adept customs officers. An application of this nature is practical even to experts or those with some knowledge of Southeast Asian taxonomy since the computerised model can reduce the number of species compared against when determining the identity of a species.

If the samples can be obtained, the code made available for this project can be used to build similar applications useful in other geographical regions of intense avian biodiversity and trade, such as Brazil, India, and certain parts of Africa. In addition, this code can easily be adapted for other taxa, particularly those with high degrees (proportion of the bird obscured by occluding objects) and rates of occlusion (how many photos in the dataset have at least some occlusion of the object of interest).

3.5.6 Understanding the motivations of end-users of applications and deployment contexts

Although artificial intelligence has many applications in wildlife crime, the deployment is often limited by funding and time constraints (Joanny, 2020). Further, many applications are now outdated (Kretser et al., 2015; Thies, 2015) and only target computer literate audiences. Particularly in lower-middle-income countries, the literacy barrier may be a potential bottleneck for a range of users (Thies, 2015).

It is undoubtedly fundamental to understand at what stage (i.e., the initial identification of species, species illegality) of the identification process machine learning technology could be used to bypass. It is also necessary to ascertain whether law enforcement uses any existing species identification tools or a potential combination of field guides, dichotomous keys or memory in bird species identification procedures. Technology could have a replacement role in monitoring markets directly or after visits, whereby a collection of photos could be uploaded after a visit to attain a list of identifications. It could also be used in the validation or verification of the presence of protected or endangered species.

Customs officers may have to be confident (i.e., state that they are 90% certain) that their species identification is correct and articulate that uncertainty, potentially in a courtroom setting if prosecution ensues. Conversely, law enforcement agencies may wish to know whether a species is protected under national legislation or not. If it is the case only to identify protected species, this would require a small subset of species for training computer models, rather than needing to identify the potentially hundreds of species that could appear in wildlife markets. Conversely, for users interested in wildlife trade, it might be more of a necessity to identify all the species at a market. Other users may have a sound knowledge of Southeast Asian bird taxonomy but not specific legislation. If a species of concern is photographed, depending on the users' pre-requisites, they can be alerted via a notification from an application.

3.5.7 Data collection concerns

We found that using data augmentation and random erasing on a small dataset of images returned a high test accuracy and F1 scores for 37 species of bird popular in the wildlife trade. This contrasts with guidelines provided in the camera trap literature which suggest that at least 1000+ label images are required per species classification of interest (Schneider et al., 2020).

To improve the data collection process in the future, we will attempt to record the bird's location from future photo submissions, if available. Consistent location metadata may speed up the manual identification process when building models of this kind and helps us understand where species are sold geographically. However, some of our photos were downloaded from open-access search engines, such as Google, or public Facebook groups, where location data and other associated metadata is difficult to ascertain. Further, as previously mentioned, if a bird is photographed in a market, it does not necessarily mean it occurs in that region. For example, Chng and Eaton (2016) found that species from eastern Indonesia were more frequently seen in markets in Java over time. Thus the utility of location metadata needs to be further explored.

Data availability is a central bottleneck for an image-based machine learning application. Although we have demonstrated proof of concept, TOT_SP_37 did not contain many sympatric species. Goliaff and Hodges (2018) recommend that researchers using wildlife images consult multiple species experts to increase confidence in their image classifications of similar sympatric species. However, if more sympatric species are included, this model may take longer to build since their manual verification a priori is more challenging due to their visual similarity. The application presented here only include 37 species, whereas more than 1,000 different species have been recorded in the wild bird trade (Harris et al., 2017).

In addition, for this project, we were limited to only the species for which we could obtain sufficient data (c.a. 100 ground-truth images). In a pre-COVID world, we would have had a different sampling strategy amongst zoos and breeding centres to build a sufficient library for more species affected by the Asian Songbird Crisis. However, during the pandemic, contact with animals in zoos was restricted even for some zoo personnel. On the other hand, only collecting data in zoos may have resulted in non-independence issues because the photos would have been taken with the same camera and very similar lighting.

Unfortunately, in the context of wildlife trade-focussed image classification applications, crowdsourcing from the public is not a feasible image labelling or processing method. This is due to the sensitive nature of wildlife trade data and the risks of providing open-source images of popular or critically endangered species. For example, Yang and Chan (2015), in their description of new gecko species in southern China, did not disclose the location of already range-restricted species due to "recurring cases of scientific descriptions driving herpetofauna to near-extinction by commercial collectors". Uploading images of birds or making bird identification technology available to the public could result in trends of species decline by commercial collectors either in the wild or in markets. Although data could be made available on request to interested parties, this may require a more careful vetting process.

In their survey on Jakarta's bird markets, Chng et al. (2015) noted that traders were fully aware that at least part of their trade is illegal. In some markets, traders did not allow photographs to be taken, and no attempts to do so were made during Chng et al.'s survey. This unofficial rule, enforced by some traders, suggests that it may be too overt to collect images in some markets, and there may be a trade-off between angering traders and collecting enough data. However, it was unclear whether this rule applied to members of the public, to law enforcement, or both. It may be the case that traders comply with law enforcement when asked, and thus they would be permitted to take photos of birds on offer.

Although in machine learning, models generally perform better when trained on more data (Schneider et al., 2020), there have been recent advances in the field of one-shot learning. One-shot learning is a novel solution to the limited data availability and has also been commonly used in facial recognition applications (Shorten & Khoshgoftaar, 2019). Atapour-Abarghouei, Bonner and McGough (2019) trained their model on a dataset of splash screens from 50 variants of ransomware (a type of malware designed to block access to a computer system), with a single image of a splash screen variant being available for each of the classes. With only ten images per class in their test dataset, they achieved test accuracies of 0.716. This approach may benefit rare bird species that are difficult to collect data for since the network only needs to see a single original image (one or a few instances) for each class. Future work can explore the application of one-shot learning to identify birds in the wildlife trade.

3.5.8 The potential of multilabel classification

Even with data bottlenecks, there are still vast opportunities to monitor markets with machine learning applications, storing invaluable metadata. In the application discussed here, each image is encoded to a singular label, its species identification. However, within TensorFlow, images can have multiple labels. Images can be tagged with multiple variables to accompany classification results, including information such as a species range, life-history traits and its CITES listing. This information can be harvested from open-source databases such as BirdLife Datazone and the Species+ database and paywall databases such as Birds of the World. In addition, given the increased importance of zoonotic disease detection, a literature review can be performed to identify zoonotic diseases associated with specific species. Birds have been known to serve as reservoirs for West Nile virus, avian influenza, salmonellosis and psittacosis (Michel et al., 2020). Subsequently, any diseases linked to a species can be included in a multilabel classification system. A disease status label will provide valuable screening capabilities to customs agencies, such as, ‘Asian Pied Starling, protected in Indonesia, not known to carry avian flu’.

3.5.9 Towards market methodologies in the future

Recording confidence is a helpful metric to be recorded in live wildlife market surveys to generate more information on baselines of human-based bird identification in the wildlife trade. There may be a general lack of self-confidence in humans in identifying tasks, not just amongst non-experts, as in Chapter 2, but also with experts. A general lack of self-confidence among closely related species, was observed by Chizinski, Martin, and Pope (2014) for angler species. When carrying out market surveys, it would be useful to have classifications accompanied by a confidence metric such as the Likert scale used in Chapter 2.

It is expected that this proposed tool will be helpful in the future, as live wildlife markets have started to reopen following a period of sustained lockdowns in Southeast Asia. Markets may be a source for zoonotic transmission to humans, particularly for avian flu (Amonsin et al., 2008; Brooks-Moizer et al., 2009; Edmunds et al., 2011) and economically important species such as domestic poultry. Using a machine learning application, of the nature presented here, may protect the health of law enforcement officers by reducing the amount of direct contact with wild specimens. Conversely, offline, mobile-based applications will facilitate covert studies in wildlife markets. Although this relies on a model with more classes and significantly more training data than the model presented here, the proof-of-concept provided in this thesis demonstrates that such a model will be feasible to build in the future, with a longer time frame for data collection and more funding.

In a COVID or post-COVID world, markets that sell birds may remain open since birds are very unlikely to be a reservoir for the SARS-CoV-2 virus or a significant transmitter. The SARS-COV-2 virus is a member of the betacoronavirus genus, whereas birds are the natural hosts of delta and gamma coronaviruses (Mahdy et al., 2020). However, the Covid-19 pandemic has also influenced bird trade by being a major driver of the online trade in wildlife (Armstrong & Chng, 2020).

Applications of this nature could be used to monitor online trade, on forums such as Facebook, or using a tool such as the MegaDetector to crop photos of wildlife from screenshots of sale items, along with web scraping tools. Hernandez-Castro and Roberts (2015) argue that technological solutions relying on images cannot be employed on other wildlife illegal markets, such as Instagram, where pictures can be missing or misleading. However, for live bird trade images are likely not misleading since each species falls in a discrete class and generally involves a clear view of the whole bird.

4.1 Conclusions

Relying on human identification capabilities only in wildlife trade monitoring is error-prone, costly, time ineffective and may hinder research capacity when automated technologies are becoming increasingly available. By using machine learning approaches, we can try and minimise some of these shortcomings. We have shown that at least some of them (consistent accuracy over time, identifying occluded images) are minimised. Given that no super-recognisers could be explicitly identified amongst a cohort of participants, analogous to customs officers, it may be the case that super-recognisers are even rarer for birds than they are for face recognition.

Customisable models, trained using transfer learning, can be modified to reflect diverse user preferences, needs and study taxa. However, further work is needed to elevate the work presented here to be deployment-ready and explore the appetite amongst law enforcement and conservation practitioners for such a specific tool as the one presented here. We also advocate for end-users voices to be included alongside the developer in human-in-the-loop workflows, which is not always the case in application development.

Despite consternation amongst members of the conservation technology community regarding building models with ‘bad data’, here we have demonstrated that image-based machine learning algorithms can be successfully applied to new, challenging domains, such as wildlife markets. It is hoped that this work will encourage others to embrace machine learning approaches to monitor megadiverse sectors of the wildlife trade.

In the face of accelerating wildlife trade and its unpredictable nature, human-in-the-loop machine learning workflows can increase the effectiveness of monitoring wildlife markets, ensure effective prosecutions and improve species identification by untrained law enforcement. More work is urgently needed in this field to address the usability of applications, counter low literacy rates in computerised systems amongst law enforcement and improve access to valuable monitoring tools.

This would feed into much-needed research on human-computer interactions relating to wildlife trade-specific machine learning applications.

References

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, *3*, e1184. https://doi.org/10.7717/peerj.1184

Alfino, S., & Roberts, D. L. (2019). Estimating identification uncertainties in CITES 'look-alike' species. *Global Ecology and Conservation*, *18*, e00648. https://doi.org/10.1016/j.gecco.2019.e00648

Aloysius, S. L. M., Yong, D. L., Lee, J. G., & Jain, A. (2020). Flying into extinction: Understanding the role of Singapore's international parrot trade in growing domestic demand. *Bird Conservation International*, *30*(1), 139–155. https://doi.org/10.1017/S0959270919000182

Amonsin, A., Choatrakol, C., Lapkuntod, J., Tantilertcharoen, R., Thanawongnuwech, R., Suradhat, S., Suwannakarn, K., Theamboonlers, A., & Poovorawan, Y. (2008). Influenza Virus (H5N1) in Live Bird Markets and Food Markets, Thailand. *Emerging Infectious Diseases*, *14*(11), 1739–1742. https://doi.org/10.3201/eid1411.080683

Armstrong, Olivia. H., & Chng, S. C. L. (2020). Distancing the flock: Bird singing competitions fly online to avoid Covid-19. *TRAFFIC Bulletin*, *32*(2), 7.

Atapour-Abarghouei, A., Bonner, S., & McGough, A. S. (2019). A King's Ransom for Encryption: Ransomware Classification using Augmented One-Shot Learning and Bayesian Approximation. *2019 IEEE International Conference on Big Data (Big Data)*, 1601–1606. https://doi.org/10.1109/BigData47090.2019.9005540

Austen, G. E., Bindemann, M., Griffiths, R. A., & Roberts, D. L. (2016). Species identification by experts and non-experts: Comparing images from field guides. *Scientific Reports*, *6*(1), 33634. https://doi.org/10.1038/srep33634

Austen, G. E., Bindemann, M., Griffiths, R. A., & Roberts, D. L. (2018). Species identification by conservation practitioners using online images: Accuracy and agreement between experts. *PeerJ*, *6*, e4157.

Barber-Meyer, S. M. (2010). Dealing with the Clandestine Nature of Wildlife-Trade Market Surveys: Hidden Trade in Wildlife Market Surveys. *Conservation Biology*, *24*(4), 918–923. https://doi.org/10.1111/j.1523-1739.2010.01500.x

Beery, S., Morris, D., & Yang, S. (2019). Efficient Pipeline for Camera Trap Image Review. *ArXiv:1907.06772 [Cs]*. http://arxiv.org/abs/1907.06772

Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in Terra Incognita. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11220, pp. 472–489). Springer International Publishing. https://doi.org/10.1007/978-3-030-01270-0_28

Bergin, D., Chng, S. C. L., Eaton, J. A., & Shepherd, C. R. (2018). The final straw? An overview of Straw-headed Bulbul Pycnonotus zeylanicus trade in Indonesia. *Bird Conservation International*, *28*(01), 126–132. https://doi.org/10.1017/S0959270917000302

Black, S. A., Fellous, A., Yamaguchi, N., & Roberts, D. L. (2013). Examining the Extinction of the Barbary Lion and Its Implications for Felid Conservation. *PLOS ONE*, *8*(4), e60174. https://doi.org/10.1371/journal.pone.0060174

Blackburn, T. M., Su, S., & Cassey, P. (2014). A Potential Metric of the Attractiveness of Bird Song to Humans. *Ethology*, *120*(4), 305–312. https://doi.org/10.1111/eth.12211

Brooks-Moizer, F., Roberton, S. I., Edmunds, K., & Bell, D. (2009). Avian Influenza H5N1 and the Wild Bird Trade in Hanoi, Vietnam. *Ecology and Society*, *14*(1). https://doi.org/10.5751/ES-02760-140128

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks: The Official Journal of the International Neural Network Society*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Burivalova, Z., Lee, T. M., Hua, F., Lee, J. S. H., Prawiradilaga, D. M., & Wilcove, D. S. (2017). Understanding consumer preferences and demography in order to reduce the domestic trade in wild-caught birds. *Biological Conservation*, *209*, 423–431. https://doi.org/10.1016/j.biocon.2017.03.005

Buschbacher, K., Ahrens, D., Espeland, M., & Steinhage, V. (2020). Image-based species identification of wild bees using convolutional neural networks. *Ecological Informatics*, *55*, 101017. https://doi.org/10.1016/j.ecoinf.2019.101017

Bush, E. R., Baker, S. E., & Macdonald, D. W. (2014). Global Trade in Exotic Pets 2006-2012: Exotic Pet Trade. *Conservation Biology*, *28*(3), 663–676. https://doi.org/10.1111/cobi.12240

Bušina, T., Kouba, M., & Pasaribu, N. (2020). What is the reliability of visually based animal trade census outcomes? A case study involving the market monitoring of the Sumatran Laughingthrush Garrulax bicolor. *Bird Conservation International*, 1–11. https://doi.org/10.1017/S095927092000026X

Butchart, S. H. M. (2008). Red List Indices to measure the sustainability of species use and impacts of invasive alien species. *Bird Conservation International*, *18*(S1), S245–S262. https://doi.org/10.1017/S095927090800035X

Cardador, L., Lattuada, M., Strubbe, D., Tella, J. L., Reino, L., Figueira, R., & Carrete, M. (2017). Regional Bans on Wild-Bird Trade Modify Invasion Risks at a Global Scale: Trade bans and invasion risk. *Conservation Letters*, *10*(6), 717–725. https://doi.org/10.1111/conl.12361

Carrete, M., & Tella, J. (2008). Wild-bird trade and exotic invasions: A new link of conservation concern? *Frontiers in Ecology and the Environment*, *6*(4), 207–211. https://doi.org/10.1890/070075

Cassey, P., Blackburn, T. M., Russell, G. J., Jones, K. E., & Lockwood, J. L. (2004). Influences on the transport and establishment of exotic bird species: An analysis of the parrots (Psittaciformes) of the world. *Global Change Biology*, *10*(4), 417–426. https://doi.org/10.1111/j.1529-8817.2003.00748.x

Challender, D. W. S., Brockington, D., Hinsley, A., Hoffmann, M., Kolby, J. E., Massé, F., Natusch, D. J. D., Oldfield, T. E. E., Outhwaite, W., Sas-Rolfes, M. 't, & Milner-Gulland, E. J. (2021). Mischaracterizing wildlife trade and its impacts may mislead policy processes. *Conservation Letters*, e12832. https://doi.org/10.1111/conl.12832

Chandler, B., & Mingolla, E. (2016). *Mitigation of Effects of Occlusion on Object Recognition with Deep Neural Networks through Low-Level Image Completion* [Research Article]. Computational Intelligence and Neuroscience; Hindawi. https://doi.org/10.1155/2016/6425257

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., McPherson, J., RStudio, library), jQuery F. (jQuery library and jQuery U., inst/www/shared/jquery-AUTHORS.txt), jQuery contributors (jQuery library; authors listed in, inst/www/shared/jqueryui/AUTHORS.txt), jQuery U. contributors (jQuery U. library; authors listed in, library), M. O. (Bootstrap, library), J. T. (Bootstrap, library), B. contributors (Bootstrap, Twitter, library), I. (Bootstrap, library), A. F. (html5shiv, library), S. J. (Respond js, library), S. P. (Bootstrap-datepicker, library), A. R. (Bootstrap-datepicker, font), D. G. (Font-A., … R), R. C. T. (tar implementation from. (2020). *shiny: Web Application Framework for R* (1.4.0.2) [Computer software]. https://CRAN.R-project.org/package=shiny

Chizinski, C. J., Martin, D. R., & Pope, K. L. (2014). Self-confidence of anglers in identification of freshwater sport fish. *Fisheries Management and Ecology*, *21*(6), 448–453. https://doi.org/10.1111/fme.12094

Chng, S. C. L., & Eaton, J. A. (2016). *In the market for extinction: Eastern and Central Java*. TRAFFIC International.

Chng, S. C. L., Eaton, J. A., Krishnasamy, K., Shepherd, C. R., & Nijman, V. (2015). *In the market for extinction: An inventory of Jakarta's bird markets.* (p. 40). TRAFFIC International.

Chng, S. C. L., Krishnasamy, K., & Eaton, J. A. (2018). In the market for extinction: The cage bird trade in Bali. *Forktail*, *34*, 7.

Choe, D., Choi, E., & Kim, D. K. (2020). *The Real-Time Mobile Application for Classifying of Endangered Parrot Species Using the CNN Models Based on Transfer Learning* [Research Article]. Mobile Information Systems; Hindawi. https://doi.org/10.1155/2020/1475164

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), 1632–1644. https://doi.org/10.1111/2041-210X.13256

Coleman, J. L., Ascher, J. S., Bickford, D., Buchori, D., Cabanban, A., Chisholm, R. A., Chong, K. Y., Christie, P., Clements, G. R., dela Cruz, T. E. E., Dressler, W., Edwards, D. P., Francis, C. M., Friess, D. A., Giam, X., Gibson, L., Huang, D., Hughes, A. C., Jaafar, Z., … Carrasco, L. R. (2019). Top 100 research questions for biodiversity conservation in Southeast Asia. *Biological Conservation*, *234*, 211–220. https://doi.org/10.1016/j.biocon.2019.03.028

Conn, P. B., McClintock, B. T., Cameron, M. F., Johnson, D. S., Moreland, E. E., & Boveng, P. L. (2013). Accommodating species identification errors in transect surveys. *Ecology*, *94*(11), 2607–2618. https://doi.org/10.1890/12-2124.1

Cooney, R., & Jepson, P. (2005). The international wild bird trade: What's wrong with blanket bans? *Oryx*, *40*(01), 18. https://doi.org/10.1017/S0030605306000056

Couret, J., Moreira, D. C., Bernier, D., Loberti, A. M., Dotson, E. M., & Alvarez, M. (2020). Delimiting cryptic morphological variation among human malaria vector species using convolutional neural networks. *PLOS Neglected Tropical Diseases*, *14*(12), e0008904. https://doi.org/10.1371/journal.pntd.0008904

Dai, C., & Zhang, C. (2017). The local bird trade and its conservation impacts in the city of Guiyang, Southwest China. *Regional Environmental Change*, *17*(6), 1763–1773. https://doi.org/10.1007/s10113-017-1141-5

Das, S. D., & Kumar, A. (2018). Bird Species Classification using Transfer Learning with Multistage Training. *ArXiv:1810.04250 [Cs]*. http://arxiv.org/abs/1810.04250

Daut, E. F., Brightsmith, D. J., Mendoza, A. P., Puhakka, L., & Peterson, M. J. (2015). Illegal domestic bird trade and the role of export quotas in Peru. *Journal for Nature Conservation*, *27*, 44–53. https://doi.org/10.1016/j.jnc.2015.06.005

Deeb, A., Roy, K., & Edoh, K. (2021). *Drone-Based Face Recognition Using Deep Learning* (pp. 197–206). https://doi.org/10.1007/978-981-15-3383-9_18

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Di Minin, E., Fink, C., Hiippala, T., & Tenkanen, H. (2019). A framework for investigating illegal wildlife trade on social media with machine learning: Social Media Content Analysis. *Conservation Biology*, *33*(1), 210–213. https://doi.org/10.1111/cobi.13104

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv:2010.11929 [Cs]*. http://arxiv.org/abs/2010.11929

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

E. Middleton, S., Lavorgna, A., Neumann, G., & Whitehead, D. (2020). Information Extraction from the Long Tail: A Socio-Technical AI Approach for Criminology Investigations into the Online Illegal Plant Trade. *12th ACM Conference on Web Science*, 82–88. https://doi.org/10.1145/3394332.3402838

Eaton, J. A., Leupen, B. T. C., & Krishnasamy, K. (2017). *Songsters of Singapore: An overview of the bird species in Singapore pet shops* (p. 35). TRAFFIC International.

Eaton, J. A., Shepherd, C. R., Rheindt, F. E., Harris, J. B. C., Van Balen, S. (Bas), Wilcove, D. S., & Collar, N. J. (2015). Trade-driven extinctions and near-extinctions of avian taxa in Sundaic Indonesia. *Forktail*, *31*, 1–12.

Edmunds, K., Roberton, S. I., Few, R., Mahood, S., Bui, P. L., Hunter, P. R., & Bell, D. J. (2011). Investigating Vietnam's Ornamental Bird Trade: Implications for Transmission of Zoonoses. *EcoHealth*, *8*(1), 63–75. https://doi.org/10.1007/s10393-011-0691-0

Farnsworth, E. J., Chu, M., Kress, W. J., Neill, A. K., Best, J. H., Pickering, J., Stevenson, R. D., Courtney, G. W., VanDyk, J. K., & Ellison, A. M. (2013). Next-Generation Field Guides. *BioScience*, *63*(11), 891–899. https://doi.org/10.1525/bio.2013.63.11.8

Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R., & Doutrelant, C. (2020). Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, *11*(9), 1072–1085. https://doi.org/10.1111/2041-210X.13436

Filter, J. (2020). *Split-folders* (0.4.3) [Python]. https://pypi.org/project/split-folders/

Fink, C., Toivonen, T., Correia, R. A., & Minin, E. D. (2021). *Mapping the online songbird trade in Indonesia*. SocArXiv. https://doi.org/10.31235/osf.io/mxkgq

Gibbon, G. E. M., Bindemann, M., & Roberts, D. L. (2015). Factors affecting the identification of individual mountain bongo antelope. *PeerJ*, *3*, e1303. https://doi.org/10.7717/peerj.1303

Gibson, K. J., Streich, M. K., Topping, T. S., & Stunz, G. W. (2019). Utility of citizen science data: A case study in land-based shark fishing. *PLOS ONE*, *14*(12), e0226782. https://doi.org/10.1371/journal.pone.0226782

Gilbert, M., Sokha, C., Joyner, P. H., Thomson, R. L., & Poole, C. (2012). Characterizing the trade of wild birds for merit release in Phnom Penh, Cambodia and associated risks to health and ecology. *Biological Conservation*, *153*, 10–16. https://doi.org/10.1016/j.biocon.2012.04.024

GIMP (2.10.22). (2020). [Computer software]. The GIMP Development Team. https://www.gimp.org/

Gomez-Villa, A., Diez, G., Salazar, A., & Diaz, A. (2016). Animal Identification in Low Quality Camera-Trap Images Using Very Deep Convolutional Neural Networks and Confidence Thresholds. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, & T. Isenberg (Eds.), *Advances in Visual Computing* (pp. 747–756). Springer International Publishing. https://doi.org/10.1007/978-3-319-50835-1_67

Gomez-Villa, A., Salazar, A., & Vargas, F. (2017). Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, *41*, 24–32. https://doi.org/10.1016/j.ecoinf.2017.07.004

Gooliaff, T. J., & Hodges, K. E. (2018). Measuring agreement among experts in classifying camera images of similar species. *Ecology and Evolution*, *8*(22), 11009–11021. https://doi.org/10.1002/ece3.4567

Harris, J. B. C., Green, J. M. H., Prawiradilaga, D. M., Giam, X., Giyanto, Hikmatullah, D., Putra, C. A., & Wilcove, D. S. (2015). Using market data and expert opinion to identify overexploited species in the wild bird trade. *Biological Conservation*, *187*, 51–60. https://doi.org/10.1016/j.biocon.2015.04.009

Harris, J. B. C., Tingley, M. W., Hua, F., Yong, D. L., Adeney, J. M., Lee, T. M., Marthy, W., Prawiradilaga, D. M., Sekercioglu, C. H., Suyadi, Winarni, N., & Wilcove, D. S. (2017). Measuring the impact of the pet trade on Indonesian birds: Bird Declines from Pet Trade. *Conservation Biology*, *31*(2), 394–405. https://doi.org/10.1111/cobi.12729

Hernandez-Castro, J., & Roberts, D. L. (2015). Automatic detection of potentially illegal online sales of elephant ivory via data mining. *PeerJ Computer Science*, *1*, e10. https://doi.org/10.7717/peerj-cs.10

Hernández-Serna, A., & Jiménez-Segura, L. F. (2014). Automatic identification of species with neural networks. *PeerJ*, *2*, e563. https://doi.org/10.7717/peerj.563

Hughes, B., & Burghardt, T. (2017). Automated Visual Fin Identification of Individual Great White Sharks. *International Journal of Computer Vision*, *122*(3), 542–557. https://doi.org/10.1007/s11263-016-0961-y

Indraswari, K., Friedman, R. S., Noske, R., Shepherd, C. R., Biggs, D., Susilawati, C., & Wilson, C. (2020). It's in the news: Characterising Indonesia's wild bird trade network from media-reported seizure incidents. *Biological Conservation*, *243*, 108431. https://doi.org/10.1016/j.biocon.2020.108431

Iñigo-Elias, E. E., & Ramos, M. A. (1991). The psittacine trade in Mexico. In *Neotropical wildlife use and conservation* (pp. 380–392). The University of Chicago Press.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv:1502.03167*, 9.

Jepson, P. (2010). Towards an Indonesian Bird Conservation Ethos: Reflections from a Study of Bird-keeping in the Cities of Java and Bali. In *Ethno-ornithology*. Routledge.

Joanny, L. (2020). *Eyes on Earth: Ensuring law enforcement technologies contribute to sustainable and just conservation* (BIOSEC: Biodiversity and Security, p. 3). University of Sheffield. https://biosec.group.shef.ac.uk/wp-content/uploads/3.-LJ-Policy-Brief-FINAL.pdf

Johansson, Ö., Samelius, G., Wikberg, E., Chapron, G., Mishra, C., & Low, M. (2020). Identification errors in camera-trap studies result in systematic population overestimation. *Scientific Reports*, *10*(1), 6393. https://doi.org/10.1038/s41598-020-63367-z

Jung, A. B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.-M., Weng, C.-H., … others. (2020). *Imgaug* (0.4.0) [Computer software]. https://github.com/aleju/imgaug

Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, *6*(4), 312–315. https://doi.org/10.1016/j.icte.2020.04.010

Khalighifar, A., Brown, R. M., Goyes Vallejos, J., & Peterson, A. T. (2021). Deep learning improves acoustic biodiversity monitoring and new candidate forest frog species identification (genus Platymantis) in the Philippines. *Biodiversity and Conservation*. https://doi.org/10.1007/s10531-020-02107-1

Komsta, L. (2011). *Package 'outliers'* (0.14) [Computer software].

Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 12.

Kretser, H. E., Wong, R., Roberton, S., Pershyn, C., Huang, J., Sun, F., Kang, A., & Zahler, P. (2015). Mobile decision-tree tool technology as a means to detect wildlife crimes and build enforcement networks. *Biological Conservation*, *189*, 33–38. https://doi.org/10.1016/j.biocon.2014.08.018

Kumar, R. L., Kakarla, J., Isunuri, B. V., & Singh, M. (2021). Multi-class brain tumor classification using residual network and global average pooling. *Multimedia Tools and Applications*, *80*(9), 13429–13438. https://doi.org/10.1007/s11042-020-10335-4

Lavorgna, A., Middleton, S. E., Pickering, B., & Neumann, G. (2020). FloraGuard: Tackling the Online Illegal Trade in Endangered Plants Through a Cross-Disciplinary ICT-Enabled Methodology. *Journal of Contemporary Criminal Justice*, *36*(3), 428–450. https://doi.org/10.1177/1043986220910297

Li, L., & Jiang, Z. (2014). International Trade of CITES Listed Bird Species in China. *PLoS ONE*, *9*(2), e85012. https://doi.org/10.1371/journal.pone.0085012

Li, Z., Wang, S.-H., Fan, R.-R., Cao, G., Zhang, Y.-D., & Guo, T. (2019). Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling. *International Journal of Imaging Systems and Technology*, *29*(4), 577–583. https://doi.org/10.1002/ima.22337

Lim, B. T. M., Sadanandan, K. R., Dingle, C., Leung, Y. Y., Prawiradilaga, D. M., Irham, M., Ashari, H., Lee, J. G. H., & Rheindt, F. E. (2019). Molecular evidence suggests radical revision of species limits in the great speciator white-eye genus Zosterops. *Journal of Ornithology*, *160*(1), 1–16. https://doi.org/10.1007/s10336-018-1583-7

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2020). On the Variance of the Adaptive Learning Rate and Beyond. *ArXiv:1908.03265 [Cs, Stat]*. http://arxiv.org/abs/1908.03265

Mahdy, M. A. A., Younis, W., & Ewaida, Z. (2020). An Overview of SARS-CoV-2 and Animal Infection. *Frontiers in Veterinary Science*, *7*, 1084. https://doi.org/10.3389/fvets.2020.596391

Marshall, B. M., Freed, P., Vitt, L. J., Bernardo, P., Vogel, G., Lotzkat, S., Franzen, M., Hallermann, J., Sage, R. D., Bush, B., Duarte, M. R., Avila, L. J., Jandzik, D., Klusmeyer, B., Maryan, B., Hošek, J., & Uetz, P. (2020). An inventory of online reptile images. *Zootaxa*, *4896*(2), 251-264-251–264. https://doi.org/10.11646/zootaxa.4896.2.6

Marshall, H., Collar, N. J., Lees, A. C., Moss, A., Yuda, P., & Marsden, S. J. (2019). Spatio-temporal dynamics of consumer demand driving the Asian Songbird Crisis. *Biological Conservation*, 108237. https://doi.org/10.1016/j.biocon.2019.108237

Marshall, H., Collar, N. J., Lees, A. C., Moss, A., Yuda, P., & Marsden, S. J. (2020). Characterizing bird-keeping user-groups on Java reveals distinct behaviours, profiles and potential for change. *People and Nature*, *2*(4). https://doi.org/10.1002/pan3.10132

McKinney, W. (2020). *pandas: A Foundational Python Library for Data Analysis and Statistics* (latest) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134

Miao, Z., Gaynor, K. M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M. S., Bowie, R. C. K., Nathan, R., Yu, S. X., & Getz, W. M. (2018). *A comparison of visual features used by humans and machines to classify wildlife* [Preprint]. Ecology. https://doi.org/10.1101/450189

Michel, N. L., Whelan, C. J., & Verutes, G. M. (2020). Ecosystem services provided by Neotropical birds. *The Condor*, *122*(3). https://doi.org/10.1093/condor/duaa022

Moreto, W. D., & Lemieux, A. M. (2015). From CRAVED to CAPTURED: Introducing a Product-Based Framework to Examine Illegal Wildlife Markets. *European Journal on Criminal Policy and Research*, *21*(3), 303–320. https://doi.org/10.1007/s10610-014-9268-0

Nash, S. V. (1993). *Sold for a song: The trade in southeast Asian non-CITES birds*. TRAFFIC International.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M. (2019). EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 10.

Niemi, J., & Tanttu, J. T. (2018). Deep Learning Case Study for Automatic Bird Identification. *Applied Sciences*, *8*(11), 2089. https://doi.org/10.3390/app8112089

Nijman, V. (2010). An overview of international wildlife trade from Southeast Asia. *Biodiversity and Conservation*, *19*(4), 1101–1114. https://doi.org/10.1007/s10531-009-9758-4

Nijman, V., Listina Sari, S., Siriwat, P., Sigaud, M., & Nekaris, K. A.-I. (2017). Records of four Critically Endangered songbirds in the markets of Java suggest domestic trade is a major impediment to their conservation. *BirdingASIA*, *27*, 20–25.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, *115*(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

Nyffeler, M., Şekercioğlu, Ç. H., & Whelan, C. J. (2018). Insectivorous birds consume an estimated 400–500 million tons of prey annually. *The Science of Nature*, *105*(7), 47. https://doi.org/10.1007/s00114-018-1571-z

Olah, G., Butchart, S. H. M., Symes, A., Guzmán, I. M., Cunningham, R., Brightsmith, D. J., & Heinsohn, R. (2016). Ecological and socio-economic factors affecting extinction risk in parrots. *Biodiversity and Conservation*, *25*(2), 205–223. https://doi.org/10.1007/s10531-015-1036-z

Olden, J. D., Lawler, J. J., & Poff, N. L. (2008). Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology*, *83*(2), 171–193. https://doi.org/10.1086/587826

Pires, S. F. (2015). The Heterogeneity of Illicit Parrot Markets: An Analysis of Seven Neo-Tropical Open-Air Markets. *European Journal on Criminal Policy and Research*, *21*(1), 151–166. https://doi.org/10.1007/s10610-014-9246-6

Pires, S. F., Olah, G., Nandika, D., Agustina, D., & Heinsohn, R. (2021). What drives the illegal parrot trade? Applying a criminological model to market and seizure data in Indonesia. *Biological Conservation*, *257*, 109098. https://doi.org/10.1016/j.biocon.2021.109098

Poon, A. (2019). *ggfree: Ggplot2-style plots with just base R graphics* [R]. https://github.com/ArtPoon/ggfree

Rahimi, S., Spiess, C. R., Gupta, B., & Sahebkar, E. (2016). *Towards Utilization of Neurofuzzy Systems for Taxonomic Identification Using Psittacines as a Case Study* [Research Article]. Applied Computational Intelligence and Soft Computing; Hindawi. https://doi.org/10.1155/2016/6798905

Rentschlar, K. A., Miller, A. E., Lauck, K. S., Rodiansyah, M., Bobby, Muflihati, & Kartikawati. (2018). A Silent Morning: The Songbird Trade in Kalimantan, Indonesia. *Tropical Conservation Science*, *11*, 1940082917753909. https://doi.org/10.1177/1940082917753909

Ribeiro, J., Reino, L., Schindler, S., Strubbe, D., Vall-llosera, M., Araújo, M. B., Capinha, C., Carrete, M., Mazzoni, S., Monteiro, M., Moreira, F., Rocha, R., Tella, J. L., Vaz, A. S., Vicente, J., & Nuno, A. (2019). Trends in legal and illegal trade of wild birds: A global assessment based on expert knowledge. *Biodiversity and Conservation*, *28*(12), 3343–3369. https://doi.org/10.1007/s10531-019-01825-5

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/BF01405730

Roberts, D. L., & Hinsley, A. (2020). The Seven Forms of Challenges in the Wildlife Trade. *Tropical Conservation Science*, *13*, 1940082920947023. https://doi.org/10.1177/1940082920947023

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face Recognition by Metropolitan Police Super-Recognisers. *PLOS ONE*, *11*(2), e0150036. https://doi.org/10.1371/journal.pone.0150036

Rosner, B., Glynn, R. J., & Lee, M.-L. T. (2003). Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach. *Biometrics*, *59*(4), 1089–1098. https://doi.org/10.1111/j.0006-341X.2003.00125.x

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Sagar, H. S. S. C., Gilroy, J. J., Swinfield, T., Yong, D. L., Gemita, E., Novriyanti, N., Lee, D. C., Janra, M. N., Balmford, A., & Hua, F. (2021). *Trade-driven trapping dampens the biodiversity benefits of forest restoration in Southeast Asia* [Preprint]. Ecology. https://doi.org/10.1101/2021.08.20.457106

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/widm.1249

Sakib, F., & Burghardt, D. T. (2021). *Visual Recognition of Great Ape Behaviours in the Wild*. 4.

Sandbrook, C., Clark, D., Toivonen, T., Simlai, T., O'Donnell, S., Cobbe, J., & Adams, W. (2021). Principles for the socially responsible use of conservation monitoring technology and data. *Conservation Science and Practice*, *3*(5), e374. https://doi.org/10.1111/csp2.374

Scheffers, B. R., Oliveira, B. F., Lamb, I., & Edwards, D. P. (2019). Global wildlife trade across the tree of life. *Science*, *366*(6461), 71–76. https://doi.org/10.1126/science.aav5327

Schneider, S. (2019). Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, *10*(4), 461-470.

Schneider, S., Greenberg, S., Taylor, G. W., & Kremer, S. C. (2020). Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and Evolution*, *10*(7), 3503–3517. https://doi.org/10.1002/ece3.6147

Şekercioğlu, Ç. H., Wenny, D. G., & Whelan, C. J. (2016). Why birds matter: Bird ecosystem services promote biodiversity and human well-being. In *Why Birds Matter: Avian Ecological Functions and Ecosystem Service* (p. 24). University of Chicago Press.

Setiyani, A. D., & Ahmadi, M. A. (2020). An overview of illegal parrot trade in Maluku and North Maluku Provinces. *Forest and Society*, 48–60. https://doi.org/10.24259/fs.v4i1.7316

Severinghaus, L. L., & Chi, L. (1999). Prayer animal release in Taiwan. *Biological Conservation*, *89*(3), 301–304. https://doi.org/10.1016/S0006-3207(98)00155-4

Sheherazade, Ober, H. K., & Tsang, S. M. (2019). Contributions of bats to the local economy through durian pollination in Sulawesi, Indonesia. *Biotropica*, *51*(6), 913–922. https://doi.org/10.1111/btp.12712

Shepherd, C., & Cassey, P. (2017). *Songbird trade crisis in Southeast Asia leads to the formation of IUCN SSC Asian Songbird Trade Specialist Group (Guest Editorial)*. *5*.

Shepherd, C. R. (2006). *The bird trade in Medan, north Sumatra: An overview*. 10.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Sibley, D. A., Bevier, L. R., Patten, M. A., & Elphicl, C. S. (2006). Comment on 'Ivory-billed Woodpecker (Campephilus principalis) Persists in Continental North America'. *Science*, *311*(5767), 1555a–1555a. https://doi.org/10.1126/science.1122778

Silvertown, J. (2015). Have Ecosystem Services Been Oversold? *Trends in Ecology & Evolution*, *30*(11), 641–648. https://doi.org/10.1016/j.tree.2015.08.007

Siriwat, P., & Nijman, V. (2020). Wildlife trade shifts from brick-and-mortar markets to virtual marketplaces: A case study of birds of prey trade in Thailand. *Journal of Asia-Pacific Biodiversity*, *13*(3), 454–461. https://doi.org/10.1016/j.japb.2020.03.012

Souto, W. M. S., Torres, M. A. R., Sousa, B. F. C. F., Lima, K. G. G. C., Vieira, L. T. S., Pereira, G. A., Guzzi, A., Silva, M. V., & Pralon, B. G. N. (2017). Singing for Cages: The Use and Trade of Passeriformes as Wild Pets in an Economic Center of the Amazon—NE Brazil Route. *Tropical Conservation Science*, *10*, 194008291768989. https://doi.org/10.1177/1940082917689898

Stringham, O. C., Toomes, A., Kanishka, A. M., Mitchell, L., Heinrich, S., Ross, J. V., & Cassey, P. (2021). A guide to using the Internet to monitor and quantify the wildlife trade. *Conservation Biology*. https://doi.org/10.1111/cobi.13675

Sykes, B. R. (2017). The elephant in the room: Addressing the Asian songbird crisis. *The Elephant in the Room*, 7.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., … Miller, R. S. (2018). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13120

Thies, I. M. (2015). User Interface Design for Low-literate and Novice Users: Past, Present and Future. *Foundations and Trends in Human–Computer Interaction*, *8*(1), 1–72. https://doi.org/10.1561/1100000047

Thompson, R. M., Hall, J., Morrison, C., Palmer, N. R., & Roberts, D. L. (2021). Ethics and governance for internet-based conservation science research. *Conservation Biology*. https://doi.org/10.1111/cobi.13778

Verma, A., Asadi, A., Yang, K., Maitra, A., & Asgeirsson, H. (2019). Analyzing household charging patterns of Plug-in electric vehicles (PEVs): A data mining approach. *Computers & Industrial Engineering*, *128*, 964–973. https://doi.org/10.1016/j.cie.2018.07.043

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, *9*(11), 2216–2225. https://doi.org/10.1111/2041-210X.13075

Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, *87*(3), 533–545. https://doi.org/10.1111/1365-2656.12780

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

Wu, Z., & Cui, Y. (2021). Edge missing image inpainting with compression–decompression network in low similarity images. *Machine Vision and Applications*, *32*(1), 37. https://doi.org/10.1007/s00138-020-01151-9

Wyatt, T. (2014). Non-Human Animal Abuse and Wildlife Trade: Harm in the Fur and Falcon Trades. *Society & Animals*, *22*(2), 194–210.

Wyatt, T., Maher, J., Allen, D., Clarke, N., & Rook, D. (2021). The welfare of wildlife: An interdisciplinary analysis of harm in the legal and illegal wildlife trades and possible ways forward. *Crime, Law and Social Change*. https://doi.org/10.1007/s10611-021-09984-9

Xin, D., Ma, L., Liu, J., Macke, S., Song, S., & Parameswaran, A. (2018). Accelerating Human-in-the-loop Machine Learning: Challenges and Opportunities. *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, 1–4. https://doi.org/10.1145/3209889.3209897

Xu, Q., Li, J., Cai, M., & Mackey, T. K. (2019). Use of Machine Learning to Detect Wildlife Product Promotion and Sales on Twitter. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00028

Yang, J.-H., & Chan, B. P.-L. (2015). Two new species of the genus Goniurosaurus (Squamata: Sauria: Eublepharidae) from southern China. *Zootaxa*, *3980*(1), 67–80. https://doi.org/10.11646/zootaxa.3980.1.4

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017). Random Erasing Data Augmentation. *ArXiv:1708.04896 [Cs]*. http://arxiv.org/abs/1708.04896

Appendix I Chapter 2 Appendix

Table A1.1: The complete data for the MA_MIS_19 dataset. The family, English name, Scientific name, IUCN conservation status, and CITES listing for the nineteen birds used in the match-mismatch experiment in Chapter 2 are compared with a computer model in Chapter 4.

Family	English name	Scientific name	IUCN Red List Status	CITES-listing
Sturnidae (Starlings)	Asian Pied Starling	*Gracupica contra*	LC	NL
Muscicapidae (Chats and Old-World flycatchers)	Bluethroat	*Cyanecula svecica*	LC	NL
Chloropseidae (Leafbirds)	Blue-masked Leafbird	*Chloropsis venusta*	NT	NL
Oriolidae (Orioles and figbirds)	Black-naped Oriole	*Oriolus chinensis*	LC	NL
Sturnidae (Starlings)	Black-winged Myna	*Acridotheres melanopterus*	CR	NL
Turdidae (Thrushes	Chestnut-backed Thrush	*Geokichla dohertyi*	NT	NL
Turdidae (Thrushes)	Chestnut-capped Thrush	*Geokichla interpres*	EN	NL
Chloropseidae (Leafbirds)	Greater Green Leafbird	*Chloropsis sonnerati*	EN	NL
Sturnidae (Starlings)	Common Hill Myna	*Gracula religiosa*	LC	II (1997)
Leiotrichidae (Laughingthrushes and allies)	Chinese Hwamei	*Garrulax canorus*	LC	II (2000)
Estrildidae (Waxbills, grass finches, munias and allies)	Java Sparrow	*Lonchura oryzivora*	EN	II (1997)
Corvidae (Crows and jays)	Javan Green Magpie	*Cissa thalassina*	CR	NL
Laniidae (Shrikes)	Long-tailed Shrike	*Lanius schach*	LC	NL
Turdidae (Thrushes)	Orange-headed Thrush	*Geokichla citrina*	LC	NL
Muscicapidae (Chats and Old-World flycatchers)	Oriental Magpie-robin	*Copyschus saularis*	LC	NL
Muscicapidae (Chats and Old-World flycatchers)	Siberian Rubythroat	*Calliope calliope*	LC	NL
Leiotrichidae (Laughingthrushes and allies)	Sumatran Laughingthrush	*Garrulax bicolor*	EN	NL
Muscicapidae (Chats and Old-World flycatchers)	White-rumped Shama	*Copsychus malabaricus*	LC	NL
Columbidae (Pigeons, Doves)	Zebra dove	*Geopelia striata*	LC	NL

Appendix II Chapter 3 Appendix

Table A2.1: The complete list of species for the TOT_SP_37 dataset. The family, English name, Scientific name, IUCN conservation status, and CITES listing is provided.

Family	English name	Scientific name	No. of ground-truth images	IUCN Red List Status	CITES-listing
Sturnidae (Starlings)	Asian Pied Starling	*Gracupica contra*	146	LC	NL
Sturnidae (Starlings)	Bali Myna	*Leucospar rothschildi*	114	CR	I (1975)
Muscicapidae (Chats and Old-World flycatchers)	Bluethroat	*Cyanecula svecica*	129	LC	NL
Chloropseidae (Leafbirds)	Blue-masked Leafbird	*Chloropsis venusta*	108	NT	NL
Oriolidae (Orioles and figbirds)	Black-naped Oriole	*Oriolus chinensis*	131	LC	NL
Sturnidae (Starlings)	Black-winged Myna	*Acridotheres melanopterus*	122	CR	NL
Chloropseidae (Leafbirds)	Blue-winged Leafbird	*Chloropsis mollucensis*	173	LC	NL
Zosteropidae (White-eyes)	Chestnut-flanked White-eye	*Zosterops erythropleurus*	139	LC	NL
Leiotrichidae (Laughingthrushes and allies)	Chestnut-capped Laughingthrush	*Garrulax mitratus*	117	NT	NL
Turdidae (Thrushes)	Chestnut-capped Thrush	*Geokichla interpres*	152	EN	NL
Estrildidae (Waxbills, grass finches, munias and allies)	Chestnut Munia	*Lonchura atricapilla*	205	LC	NL
Sturnidae (Starlings)	Common Myna	*Acridotheres tristis*	164	LC	NL
Psittacidae (Parrots)	Fischer’s Lovebird	*Agapornis fischeri*	129	NT	II (2005)
Chloropseidae (Leafbirds)	Greater Green Leafbird	*Chloropsis sonnerati*	133	EN	NL
Psittacidae (Parrots)	Grey Parrot	*Psittacus erithacus*	156	EN	I (2017
Sturnidae (Starlings)	Common Hill Myna	*Gracula religiosa*	254	LC	II (1997)
Leiotrichidae (Laughingthrushes and allies)	Chinese Hwamei	*Garrulax canorus*	136	LC	II (2000)

Table A2.1: Continued.

Fringillidae (Finches and Hawaiian honeycreeepers)	Japanese Grosbeak	*Eophona personata*	228	LC	NL
Estrildidae (Waxbills, grass finches, munias and allies)	Java Sparrow	*Lonchura oryzivora*	180	EN	II (1997)
Corvidae (Crows and jays)	Javan Green Magpie	*Cissa thalassina*	117	CR	NL
Leiotrichidae (Laughingthrushes and allies)	Red-billed Leiothrix	*Leiothrix lutea*	162	LC	II (1997)
Laniidae (Shrikes)	Long-tailed Shrike	*Lanius schach*	151	LC	NL
Turdidae (Thrushes)	Orange-headed Thrush	*Geokichla citrina*	163	LC	NL
Muscicapidae (Chats and Old-World flycatchers)	Oriental Magpie-robin	*Copyschus saularis*	201	LC	NL
Leiotrichidae (Laughingthrushes and allies)	Rufous-fronted Laughingthrush	*Garrulax rufifrons*	185	CR	NL
Pycnonotidae (Bulbuls)	Red-whiskered Bulbul	*Pycnonotus jocosus*	210	LC	NL
Sturnidae (Starlings)	Red-billed Starling	*Spodiospar sericeus*	222	LC	NL
Muscicapidae (Chats and Old-World flycatchers)	Siberian Rubythroat	*Calliope calliope*	108	LC	NL
Leiotrichidae (Laughingthrushes and allies)	Sumatran Laughingthrush	*Garrulax bicolor*	123	EN	NL
Leiotrichidae (Laughingthrushes and allies)	Silver-eared Mesia	*Leiothrix argentauris*	166	LC	II (1997)
Estrildidae (Waxbills, grass finches, munias and allies)	Scaly-breasted Munia	*Lonchura punctulata*	239	LC	NL
Pycnonotidae (Bulbuls)	Straw-headed Bulbul	*Pycnonotus zeylanicus*	118	CR	II (1997)
Zosteropidae (White-eyes)	Swinhoe's White-eye	*Zosterops simplex*	229	LC	NL
Estrildidae (Waxbills, grass finches, munias and allies)	White-rumped Munia	*Lonchura striata*	226	LC	NL
Muscicapidae (Chats and Old-World flycatchers)	White-rumped Shama	*Copsychus malabaricus*	269	LC	NL
Columbidae (Pigeons, Doves)	Zebra Dove	*Geopelia striata*	198	LC	NL
Estrildidae (Waxbills, grass finches, munias and allies)	Zebra Finch	*Taenopygia castanotis*	133	LC	NL

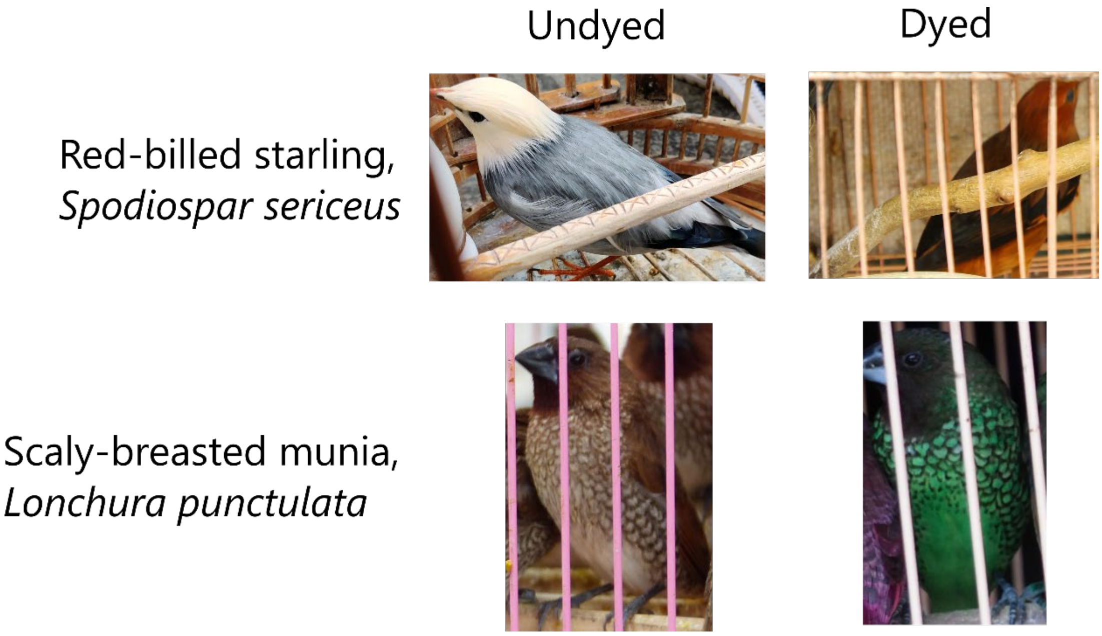


Figure A2.1: The effect of artificial dying on species appearance, as occasionally seen when Red-billed Starling's, *Spodiospar sericeus* and Scaly-breasted Munia's, *Lonchura punctulata* are offered for sale.

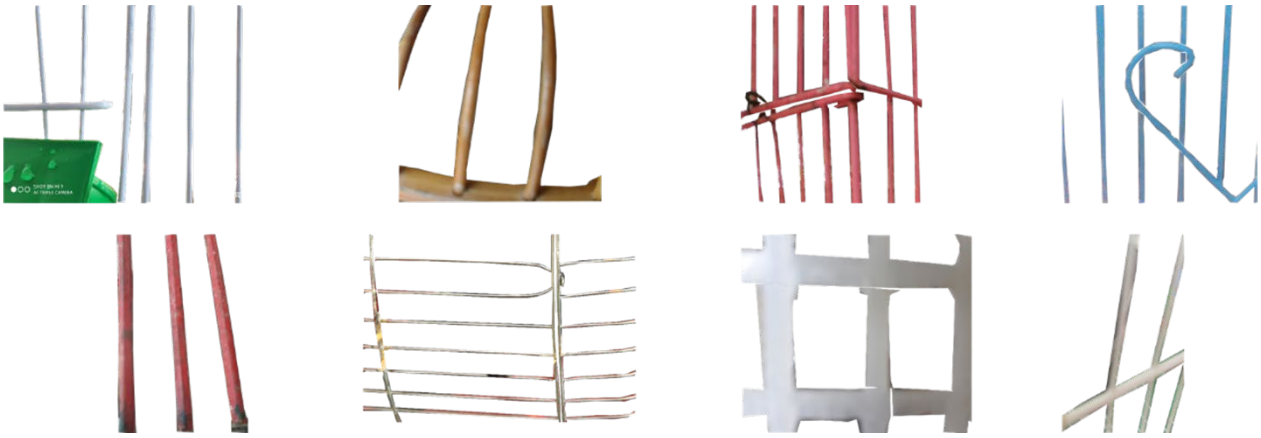


Figure A2.2: A sample of cage masks generated from original bird images with cages in the foreground in the image editing software, GIMP.

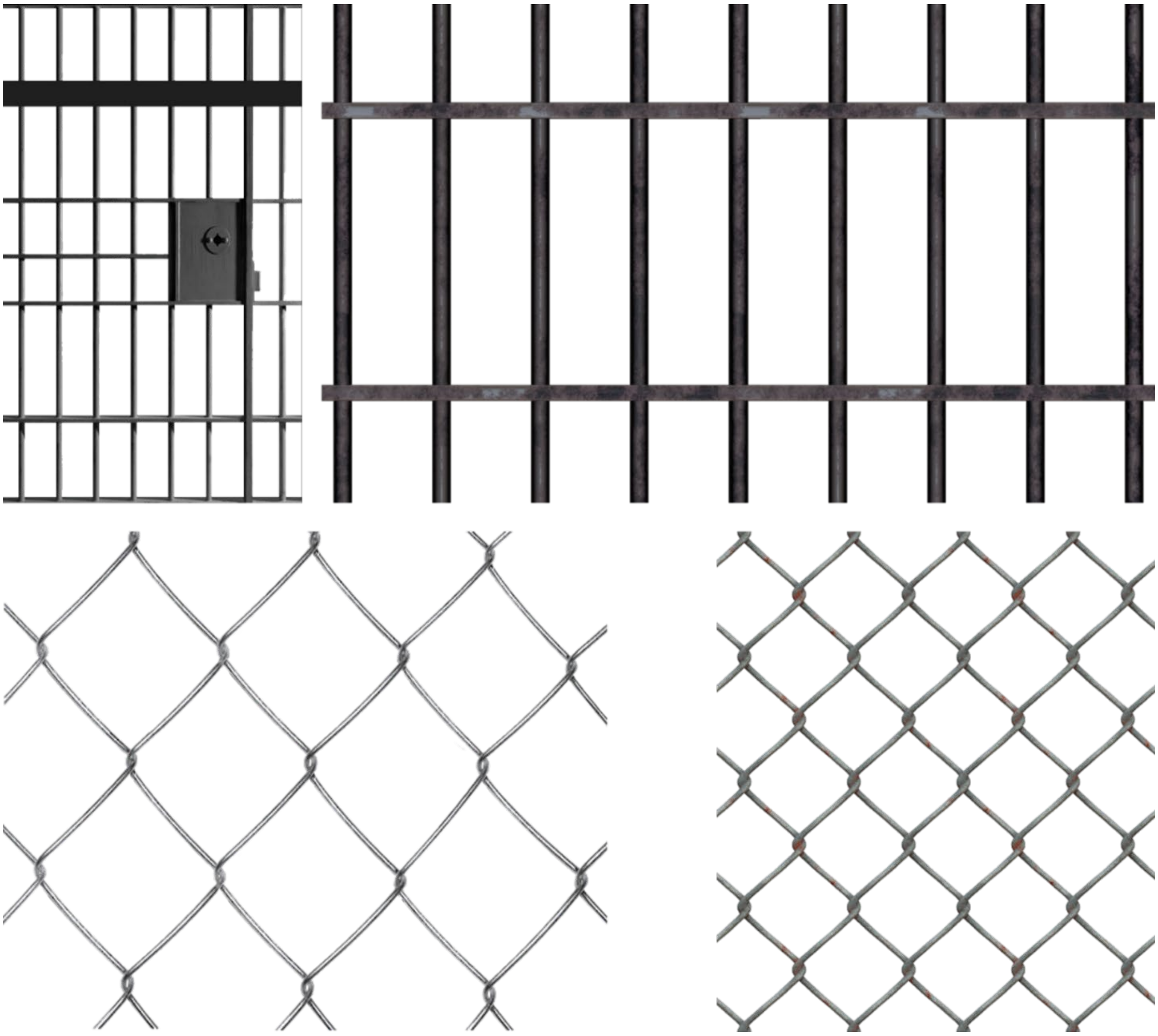


Figure A2.3: Cage bars, prison, and transparent fence pngs downloaded from Google.

Appendix II.1 Image inpainting

Image inpainting is used to minimise the appearance of damaged sections by painting over those sections in a manner consistent with the rest of the image (Nazeri et al., 2019; Wu & Cui, 2021). A large-scale application of image inpainting is the Edgeconnect model. The edge generator hallucinates edges of the missing region (both regular and irregular) of the image. The image completion network fills in the missing regions using hallucinated edges as a priori. In terms of conservation, edge detection algorithms have been applied to evaluate the relative camouflage of nesting shorebird species compared to their nesting substrate (Weinstein, 2018).

Here, we demonstrate the applicability of the model by plotting some examples, using a model pre-trained on the places2 dataset, which was created using a Graphical User Interface (GUI) of the EdgeConnect model (available at https://github.com/icepoint666/edge-connect-ui). The Places dataset contains more than 10 million images comprising 400+ unique scene categories. This dataset was chosen as some of the photos also contained birds. Using the GUI, the user uploads their image, paints the mask over the areas of occlusion, and then the output is an image with the occluded areas filled in, which can then be saved locally.

Image inpainting is an effective technique, as we can see from the figure. However, even inpainting one image with the assistance of a GUI is lengthy. On average, the time needed to fill in damaged or occluded portions of the image could be between 10 and 30 seconds per image, along with nine seconds for inference. This process is made lengthier by manually selecting the image in the GUI and saving it post-inference.

Given that we had 5,963 ground-truth images with at least some degree of occlusion, where the view of the bird was obstructed, this would have taken at least 55 hours of manual effort if the image did not need further correction. In some cases, the image inpainting did not fill in correctly, whereby the filled-in area was blurry or still contained part of a cage bar. This could be corrected by using a larger brushstroke or masking a more extensive area out. However, due to our dataset's high degree of occlusion, we still require a more scalable solution to treat occluded images.

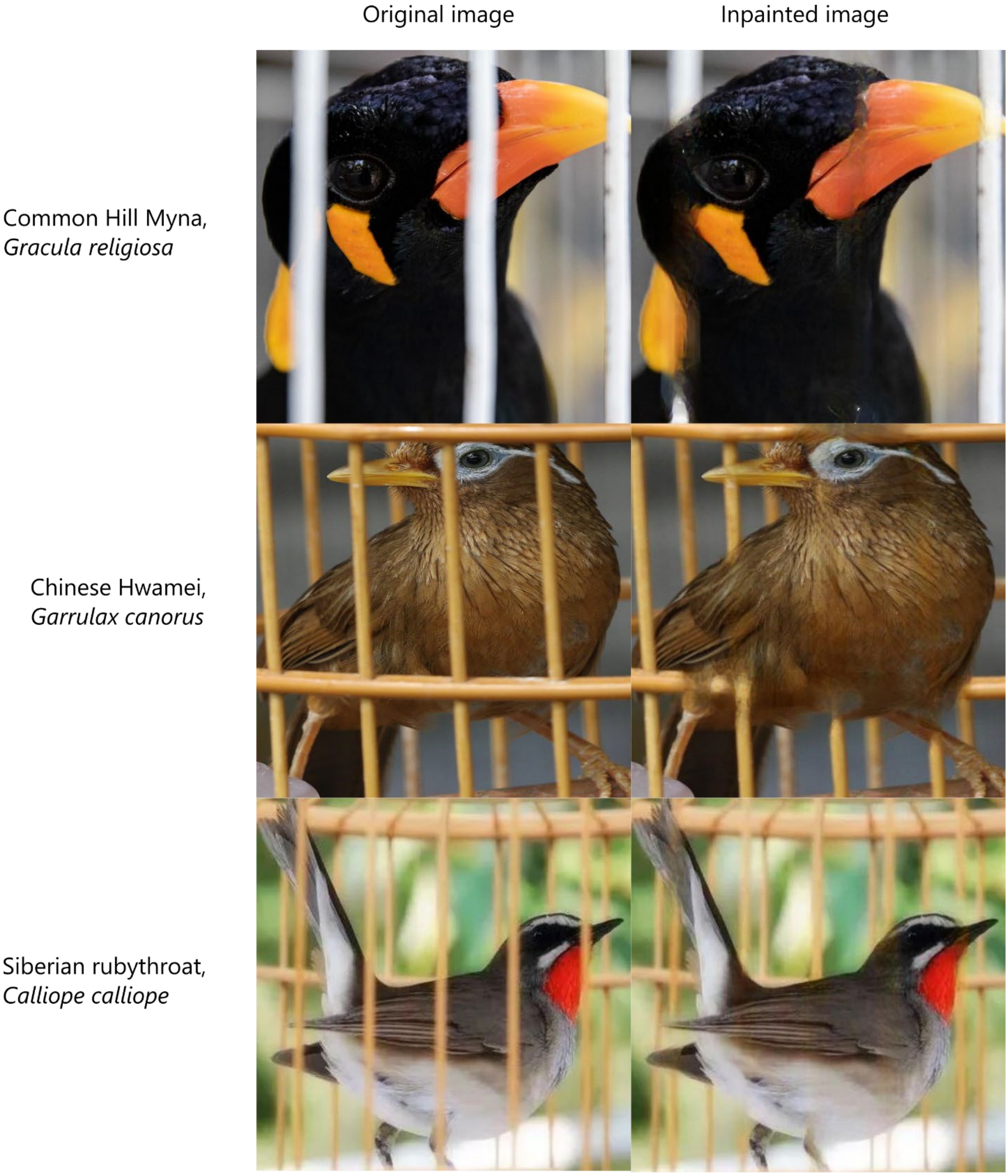


Figure A2.4: The effect of image inpainting using the EdgeConnect GUI on three different species.

Table A2.2: The classification report containing precision, recall and F1 score for each of the species in the test set of TOT_SP_37.

Species	Precision	Recall	F1 Score	Support (no. of photos)
Asian Pied Starling	0.90	1.00	0.95	38
Bali Myna	0.92	0.90	0.91	39
Bluethroat	0.91	1.00	0.95	39
Blue-masked Leafbird	0.97	0.94	0.96	35
Black-naped Oriole	0.97	0.97	0.97	40
Blue-winged Leafbird	0.97	0.87	0.92	38
Black-winged Myna	1.00	0.89	0.94	37
Chestnut-capped Laughingthrush	1.00	0.86	0.93	36
Chestnut-flanked White-eye	0.95	0.92	0.94	39
Chestnut capped Thrush	0.95	0.97	0.96	37
Chestnut Munia	0.97	0.88	0.92	32
Common Myna	0.97	0.92	0.94	37
Fischer’s Lovebird	0.95	0.95	0.95	39
Greater Green Leafbird	0.86	0.97	0.91	38
Grey Parrot	0.80	0.97	0.88	36
Japanese Grosbeak	0.97	0.84	0.90	38
Common Hill Myna	0.95	1.00	0.97	39
Chinese Hwamei	0.97	0.93	0.95	40
Java Sparrow	0.86	0.95	0.90	39
Javan Green Magpie	0.93	1.00	0.96	38
Red-billed Leiothrix	0.97	0.89	0.93	38
Long-tailed Shrike	1.00	0.92	0.96	38
Orange-headed Thrush	0.97	1.00	0.99	39
Oriental Magpie-robin	0.93	0.97	0.95	38
Rufous-fronted Laughingthrush	0.97	1.00	0.99	38
Red-billed Starling	0.94	0.81	0.87	37

Table A2.2: Continued.

Red-whiskered Bulbul	1.00	0.97	0.99	38
Siberian Rubythroat	0.89	0.94	0.91	33
Scaly-breasted Munia	0.97	1.00	0.99	37
Straw-headed Bulbul	0.97	1.00	0.99	36
Silver-eared Mesia	1.00	0.79	0.88	38
Sumatran Laughingthrush	0.97	0.97	0.97	33
Swinhoe’s White-eye	0.94	0.97	0.96	35
White-rumped Munia	0.94	0.86	0.90	35
White-rumped Shama	0.95	0.88	0.91	41
Zebra Dove	0.95	0.90	0.92	39
Zebra Finch	0.67	0.95	0.79	39

Table A2.3: The training data similarity matrix across the five folds, partitioned from the original TOT_SP_37. There are 6,110 photos in each fold.

	Fold_0 train	Fold_1 train	Fold_2 train	Fold_3 train	Fold_4 train
Fold_0 train	1	0.696	0.699	0.703	0.701
Fold_1 train	0.696	1	0.699	0.703	0.694
Fold_2 train	0.699	0.699	1	0.700	0.699
Fold_3 train	0.703	0.703	0.700	1	0.694
Fold_4 train	0.701	0.694	0.699	0.694	1

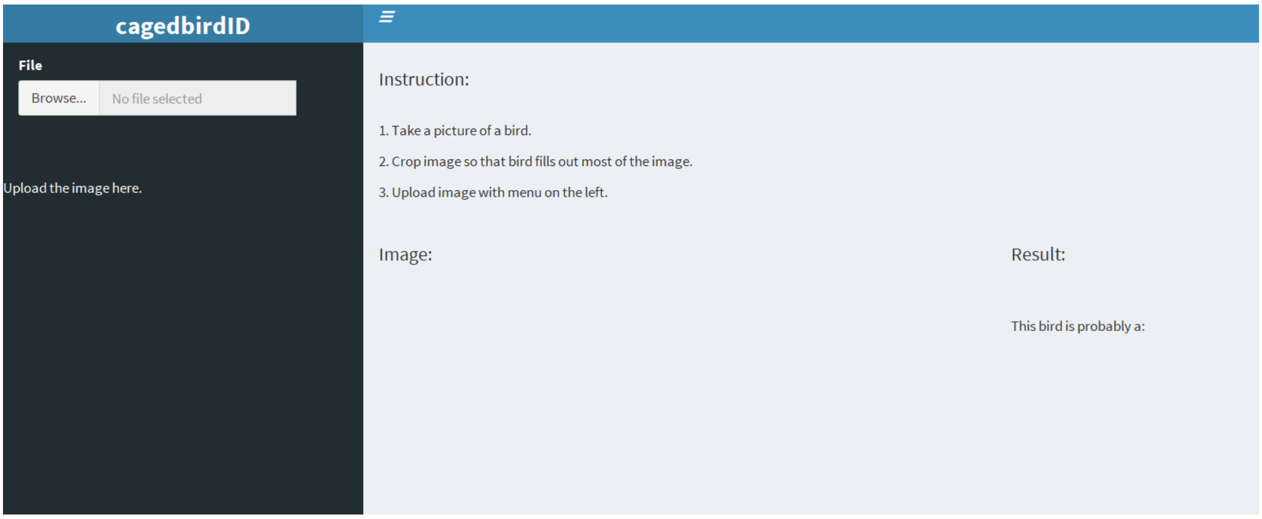
Table A2.4: The validation data similarity matrix across the five folds, partitioned from the original TOT_SP_37. There are 6,110 photos in each fold.

	Fold_0 val	Fold_1 val	Fold_2 val	Fold_3 val	Fold_4 val
Fold_0 val	1	0.137	0.153	0.142	0.168
Fold_1 val	0.137	1	0.153	0.142	0.144
Fold_2 val	0.153	0.153	1	0.137	0.147
Fold_3 val	0.142	0.142	0.137	1	0.148
Fold_4 val	0.168	0.144	0.147	0.148	1

Table A2.5: The test data similarity matrix across the five folds, partitioned from the original TOT_SP_37. There are 6,110 photos in each fold.

	Fold_0 test	Fold_1 test	Fold_2 test	Fold_3 test	Fold_4 test
Fold_0 test	1	0.170	0.153	0.168	0.157
Fold_1 test	0.170	1	0.159	0.141	0.136
Fold_2 test	0.153	0.159	1	0.157	0.145
Fold_3 test	0.168	0.141	0.157	1	0.146
Fold_4 test	0.157	0.136	0.145	0.146	1

Appendix II.2 Predictions from the Shiny app interface



Upload the image here.

Instruction:

1. Take a picture of a bird.
2. Crop image so that bird fills out most of the image.
3. Upload image with menu on the left.

Image:

Result:

This bird is probably a:

Figure A2.5: A screenshot of the GUI of the Shiny application, deployed locally.

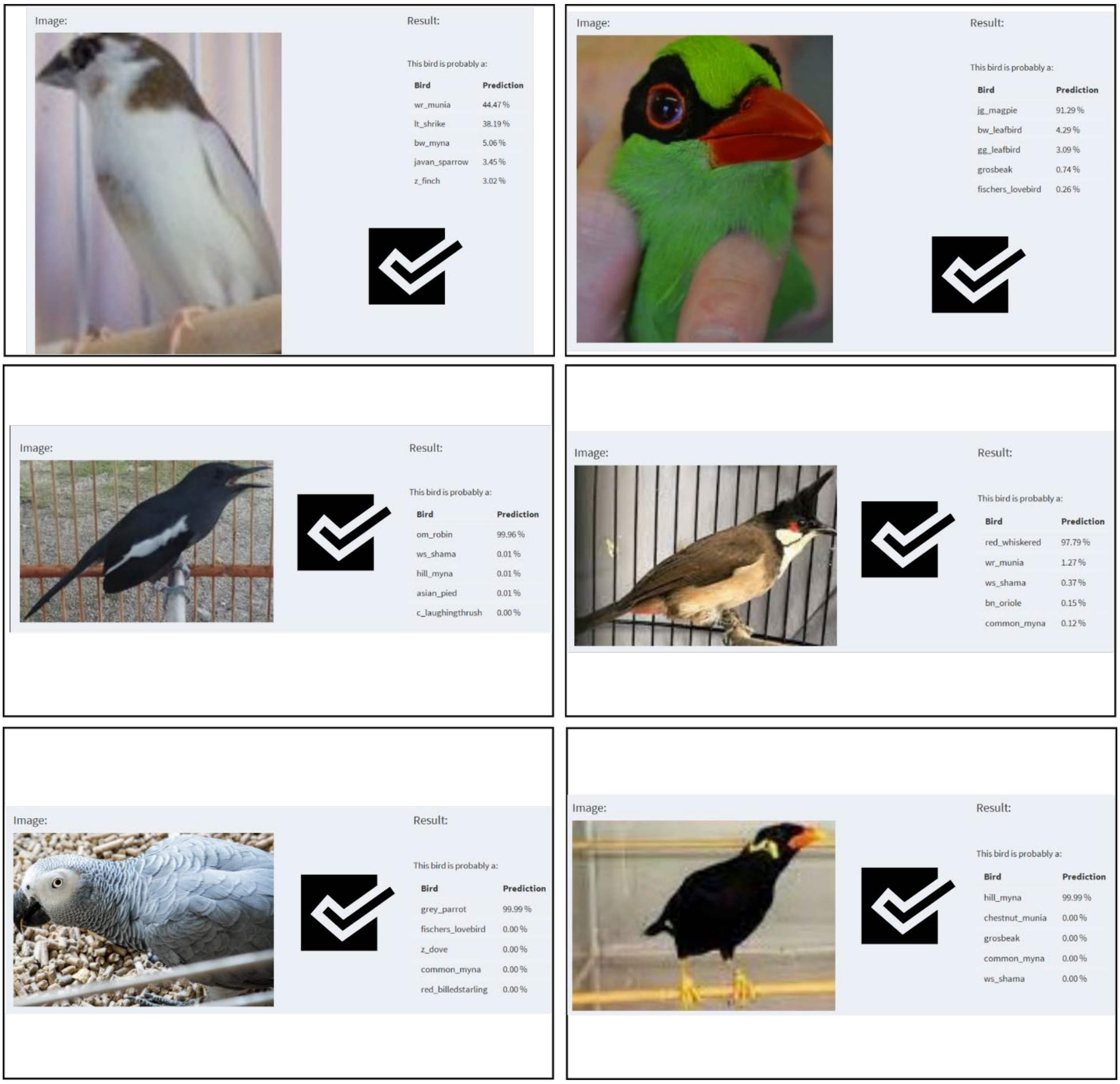


Image:

Result:

This bird is probably a:

Bird	Prediction
wr_munia	44.47 %
lt_shrike	38.19 %
bw_myna	5.06 %
javan_sparrow	3.45 %
z_finch	3.02 %

✓

Image:

Result:

This bird is probably a:

Bird	Prediction
jg_magpie	91.29 %
bw_leafbird	4.29 %
gg_leafbird	3.09 %
grosbeak	0.74 %
fischers_lovebird	0.26 %

✓

Image:

Result:

This bird is probably a:

Bird	Prediction
om_robin	99.96 %
ws_shama	0.01 %
hill_myna	0.01 %
asian_pied	0.01 %
c_laughingthrush	0.00 %

✓

Image:

Result:

This bird is probably a:

Bird	Prediction
red_whiskered	97.79 %
wr_munia	1.27 %
ws_shama	0.37 %
bn_oriole	0.15 %
common_myna	0.12 %

✓

Image:

Result:

This bird is probably a:

Bird	Prediction
grey_parrot	99.99 %
fischers_lovebird	0.00 %
z_dove	0.00 %
common_myna	0.00 %
red_billedstarling	0.00 %

✓

Image:

Result:

This bird is probably a:

Bird	Prediction
hill_myna	99.99 %
chestnut_munia	0.00 %
grosbeak	0.00 %
common_myna	0.00 %
ws_shama	0.00 %

✓

Figure A2.6: The performance of the Shiny model applied to a sample of uncaged, cropped images, along with the top-5 accuracy.

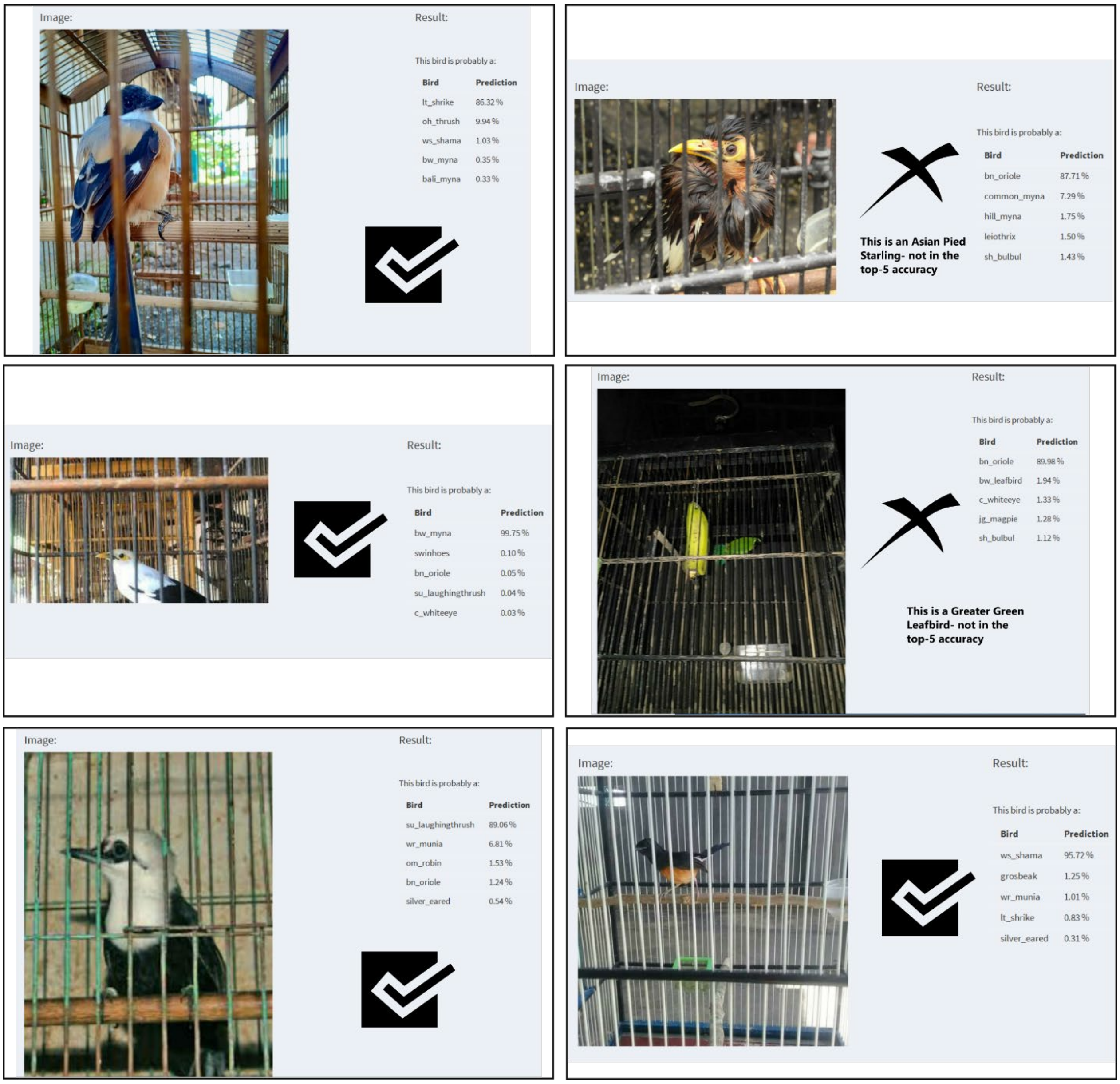


Image:

Result:

This bird is probably a:

Bird	Prediction
lt_shrike	86.32 %
oh_thrush	9.94 %
ws_shama	1.03 %
bw_myna	0.35 %
bali_myna	0.33 %

✓

Image:

Result:

✗

This is an Asian Pied Starling- not in the top-5 accuracy

This bird is probably a:

Bird	Prediction
bn_oriole	87.71 %
common_myna	7.29 %
hill_myna	1.75 %
leiothrix	1.50 %
sh_bulbul	1.43 %

Image:

Result:

✓

This bird is probably a:

Bird	Prediction
bw_myna	99.75 %
swinhoes	0.10 %
bn_oriole	0.05 %
su_laughingthrush	0.04 %
c_whiteeye	0.03 %

Image:

Result:

This bird is probably a:

Bird	Prediction
bn_oriole	89.98 %
bw_leafbird	1.94 %
c_whiteeye	1.33 %
jg_magpie	1.28 %
sh_bulbul	1.12 %

✗

This is a Greater Green Leafbird- not in the top-5 accuracy

Image:

Result:

This bird is probably a:

Bird	Prediction
su_laughingthrush	89.06 %
wr_munia	6.81 %
om_robin	1.53 %
bn_oriole	1.24 %
silver_eared	0.54 %

✓

Image:

Result:

✓

This bird is probably a:

Bird	Prediction
ws_shama	95.72 %
grosbeak	1.25 %
wr_munia	1.01 %
lt_shrike	0.83 %
silver_eared	0.31 %

Figure A2.7: The performance of the Shiny model applied to a sample of caged, uncropped images, along with the top-5 accuracy.

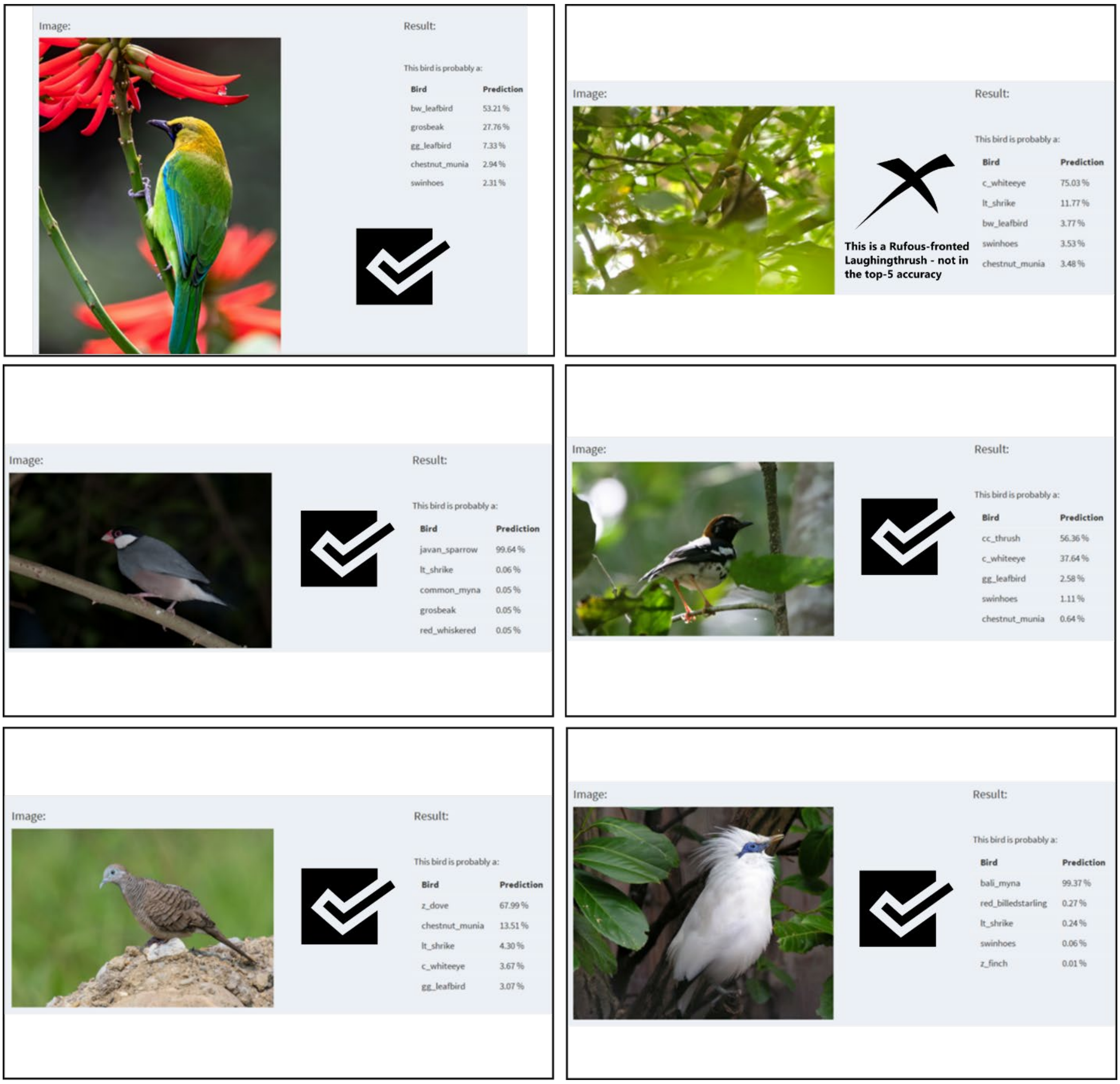


Image:

Result:

This bird is probably a:

Bird	Prediction
bw_leafbird	53.21 %
grosbeak	27.76 %
gg_leafbird	7.33 %
chestnut_munia	2.94 %
swinhoes	2.31 %

✓

Image:

Result:

This is a Rufous-fronted Laughingthrush - not in the top-5 accuracy

This bird is probably a:

Bird	Prediction
c_whiteeye	75.03 %
lt_shrike	11.77 %
bw_leafbird	3.77 %
swinhoes	3.53 %
chestnut_munia	3.48 %

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
javan_sparrow	99.64 %
lt_shrike	0.06 %
common_myna	0.05 %
grosbeak	0.05 %
red_whiskered	0.05 %

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
cc_thrush	56.36 %
c_whiteeye	37.64 %
gg_leafbird	2.58 %
swinhoes	1.11 %
chestnut_munia	0.64 %

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
z_dove	67.99 %
chestnut_munia	13.51 %
lt_shrike	4.30 %
c_whiteeye	3.67 %
gg_leafbird	3.07 %

Image:

✓

Result:

This bird is probably a:

Bird	Prediction
bali_myna	99.37 %
red_billedstarling	0.27 %
lt_shrike	0.24 %
swinhoes	0.06 %
z_finch	0.01 %

Figure A2.8: The performance of the Shiny model applied to a sample of wild, uncropped images, along with the top-5 accuracy.