Interpretable Wildlife Classification by Coupling Genetics, Scoring Systems, and Computer Vision

Sara Gonzalez[a]*, Andrew Lensen[b] and Philip Lavretsky[a]

[a]*Department of Biology, University of Texas at El Paso. 500 W. University, El Paso, TX 79968, United States;*

[b]*School of Engineering and Computer Science, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand*

***Corresponding Author:**
Sara Gonzalez
Department of Biology
University of Texas at El Paso
500 W. University Ave
El Paso, TX 79968
United States
sgonzalez28@miners.utep.edu

Abstract

Despite increasing interest in using machine learning to improve field-based species identification, key challenges remain, particularly limited model interpretability and the large datasets typically required for reliable performance. Additionally, model accuracy often depends on expert-labelled training data, which may reflect the same assumptions about trait diagnosability that constrain traditional methods. We develop interpretable, trait-specific machine learning protocols designed to perform effectively with relatively small datasets. We incorporate heatmaps to assess whether poor model performance arises from data limitations, modeling artifacts, or irrelevant image features. Interpretability is further enhanced through a numerical scoring system applied to individual diagnostic traits, with final classifications based on aggregated trait values rather than a single model output. Using transfer learning with a pretrained ResNet34 architecture, we developed ten neural network models trained on images representing diagnostic phenotypic traits. Models were trained on photographs of genetically characterized individuals, reducing reliance on expert identification alone. We applied this framework to distinguish between two closely related and historically difficult-to-identify species, the Mexican Duck (*Anas diazi*) and the Mallard (*Anas platyrhynchos*). Trait-based models achieved balanced accuracies exceeding 90%, accounting for an unbalanced training dataset. However, when evaluated against expert and non-expert classifications for the same 30 individuals, overall AI performance more closely resembled that of non-experts. Heatmap analyses revealed inconsistent attention to background features, indicating sensitivity to image artifacts. This study demonstrates how integrating genomics, interpretable AI, and multi-trait analyses can reduce bias in phenotypic identification and improve classification performance with limited datasets.

Keywords: Artificial Intelligence; Computer Vision; Convolutional Neural Networks; Binary Classification; Image Segmentation; Wildlife Conservation; Waterfowl

1. Introduction

Application of artificial intelligence (AI) in biology has become more accessible by harnessing the growing power of computer vision and machine learning with technological advances in Graphic Processing Units (GPUs) (Liu & Du, 2024). Generally, machine learning is the process by which computers conduct pattern recognition through specified algorithms on provided data and is a tool that has been widely used throughout research practices, including towards species identification (Nadimpalli, 2005, Brenowitz & Larson, 2015, Atanbori et al, 2016). For wildlife, machine learning has been used to identify individuals from data ranging from sound pattern recognition (Brandes, 2008; Incze et al, 2018; Mehyadin et al, 2021) to color histograms (Marini et al, 2013; Miao et al, 2019; Moallem, 2021). These studies demonstrate the ability to not only automate the process but help avoid individual-based biases that often arise (Mendes et al, 2024). However, machine learning still requires sufficient and high-quality data vetted by experts in the field (Picard et al, 2020; Gong et al, 2023). While such a model framework has been well integrated into large species databases like Merlin Bird (i.e., ebird; Chu, 2012). and iNaturalist (Ueda & Loarie, 2008), these remain contingent on those expert's initial identification and constant improvements based on the processing of hundreds of thousands of images constantly being fed by end users. Consequently, such applications are not often feasible for most researchers and their study systems (Farley et al, 2018; Hampton et al, 2013). Furthermore, the identification of phenotypically similar species can be highly complex even for the most trained individuals (Müller et al, 2007), which becomes even more convoluted if such species make viable hybrid offspring in the wild. Here, we attempt to circumvent such limitations by integrating methods and processes with datasets of genetically rather than expertise verified individuals into a machine learning algorithm. In addition to understanding the potential for

identifying parental and hybrid individuals, we assess any limitations of such pipelines for datasets of moderate sizes that are more realistic for most wildlife systems. Importantly, unlike many previous approaches that rely on widely available public image sets (Man & Chahl, 2022), this study employs a newly assembled training dataset in which all images are molecularly verified. Doing so has the potential to remove the need for trained experts to verify and remove potentially unknown biases when solely dealing with phenotypic traits (Adolph & Hardin, 2007; Van der Sluis et al, 2010). In addition, the training dataset captures a broad range of phenotypic variation within each species, minimizing the risk that unaccounted plumage variation biases model performance. Although AI has been used to study genic-disease correlations in human diseases (Duong et al, 2022; Reyes & Sanchez, 2024), this is the first attempt to determine trait-genetic associations from pictures using artificial intelligence and computer vision.

Here, we aim to develop methods that standardize species identification with measurable confidence for non-experts, while accounting for data limitations. Critically, this study contributes to both ecological and AI research by emphasizing model interpretability by designing architectures that mimic expert decision-making processes and applying visual explainability tools such as activation heat maps to reveal regions influencing model predictions within images. Together, these efforts move beyond simple classification accuracy toward a biologically meaningful understanding of how models "see," ensuring that automated systems make ecologically valid classifications rather than relying on spurious cues. By integrating interpretability and biological reasoning, this work advances the development of transparent, reliable, and scalable AI approaches for ecological trait detection.

To build interpretable classification models, we deployed convolutional neural networks (CNNs), which identify patterns in images through layered filtering operations (LeCun et al.,

2015). Specifically, we implemented a CNN architecture based on a ResNet34 backbone trained through transfer learning, allowing the model to leverage pre-trained visual features while fine-tuning to recognize biologically relevant plumage traits. Although deeper networks tend to yield more robust performance, their increasing complexity also reduces interpretability, producing hyperdimensional representations that are difficult to link to human logic an ultimately lead to the "black box" effect (Murdoch et al., 2019). To mitigate this, we constrained the model architecture to mimic the process biologists use when identifying species by evaluating the presence or absence of individual diagnostic traits within specialized models. This approach allows causes of model decisions to be extracted more transparently. we further enhanced interpretability using heat maps of the final convolutional layer to visualize which image regions drive classification decisions (Lipton, 2018; Lucas, 2020), enabling an assessment of whether the model relied on biological relevant cues or on spurious image artifacts. Although constraining the model in this way may reduce its maximal predictive power, we considered this an appropriate tradeoff as it substantially improves interpretability and allows us to evaluate the biological validity of the model's decisions.

We applied these models to a case study involving two closely related duck species, the Mexican Duck (*Anas diazi*) and Mallard (*Anas platyrhynchos*). Both belong to the Mallard Complex, a radiation of 14 species that diverged over the last two million years (Lavretsky, 2021), with Mexican Ducks and Mallards representing the most recent split (~500,000 years; Lavretsky et al., 2015; Brown et al., 2022a), and population genomic studies have repeatedly shown that many phenotypic traits used in traditional field keys fail to reliably distinguish parental individuals from hybrids (Lavretsky et al, 2019; Lavretsky et al., 2021; Brown et al., 2022b). For example, keys designed by Kirby et al. (2000) correctly classified only 80% of parental birds and

60% of hybrids (Lavretsky et al., 2019). Because misidentification can introduce substantial bias into management metrics such as sex-age structure estimates (McCartney et al., 2019; Perry et al., 2002), improved diagnostic approaches are needed. Recent genomic-first trait analyses have demonstrated >98% accuracy when distinguishing sex-age cohorts of Mexican Ducks, Mallards and their hybrids (Brown et al., 2022b). We build directly on this work by training models using photographs of genetically vetted individuals, enabling classification of parental species with hybrids without reliance on phenotype-driven expert judgement. This approach also allows us to evaluate the capabilities and limitations of model AI-based methods independently of observer bias and to determine whether classification challenges arise from phenotypic similarity, data quality, or algorithmic constraints.

2. Methods

A total of 11 phenotypic traits were considered across images (Table 1). Although initial testing used whole-bird photographs, we found that the potential of computer vision to score traits increased by partitioning each trait by body region including, the dorsal and ventral sides of the head, body, and wings (Figure 1). This ensured that the neural network was trained on images that are directly relevant to each trait, reducing the likelihood of misclassification due to irrelevant plumage or background features. For example, classification accuracy for detecting green coloration on the head increased markedly when using head-specific partitions rather than full-body images (Figure 2).

Once corresponding trait folders were prepared, we performed data augmentation using the Python Albumentations package (Buslaev et al, 2020), applying randomized rotation, transpositions, crops, and small perturbations in hue, saturation and value while remaining within the RGB color space. Alternate color spaces (e.g. HSV) were not used at this stage due as RGB is

the native color format of the photographic dataset and is the default input representation for ResnNet-based architectures, and thus, models were trained directly on RGB tensors. This avoided unnecessary color-space transformations while maintaining compatibility with established computer-vision pipelines. These alterations increased the effective size of training sets. Moreover, using models like ResNet34 (He et al, 2015) that is pretrained on approximately 1.3 million ImageNet images reduced the total amount of training required through transfer learning. During training, we monitored performance using trained and validation loss, error rate, balanced accuracy and F1 score. Balanced accuracy was used due to an uneven distribution in quantity of images between Mexican Ducks (N = 2,467) and Mallards (N = 834) (Brodersen et al, 2010), while the F1 score used to determine accuracy using precision and recall in machine learning (Sørensen, 1948).

After training, transfer learning was done using the *fast.ai* package in which the full data pipeline takes the augmented images and separates 20% of them for validation. Following *fast.ai*'s protocol, a learning rate was determined for each learning run (Howard & Gugger 2020), which ensured that the learning rate wasn't too slow or fast that could cause the model to take too long to learn or the model to overshoot and likewise fail to train properly, respectively. The model is first trained for 50 cycles (called 'epochs') using the 1Cycle Policy (Smith & Topin, 2019). The 1Cycle Policy is a method designed to accelerate neural network training by dynamically adjusting the learning rate. An early stopping criterion was applied to halt training if no improvement in training or validation loss was observed for three consecutive epochs, helping to prevent overfitting. During this phase, all layers of the neural network except the last two were frozen, allowing only these last layers to be updated. This approach leverages the fact that earlier convolutional layers typically capture low-level features such as edges and textures, while deeper

layers encode more complex visual patterns. These weights were saved as "st1" (step one) once the training completed. Next, the model is completely unfrozen, and a new learning rate is found; once again, the network is trained for 50 epochs (with the same early-stop criterion) with the new learning rate though the rate is severely reduced for the topmost layers as these are already trained on finding basic shapes and patterns that do not need to be re-learned. This second pass is saved separately as "unfrozen." Note all model training was done with a RTX 4000 ADA Generation Graphics Processing Unit (GPU), on a Windows Subsystem for Linux (WSL) using an Ubuntu 24.04.2 distribution. Our pipeline also readies the transfer learning architecture to have the desired amount of output nodes at the final layer.

With a binary classification system, our final layers had two possible outputs as either Mexican Duck or Mallard. In short, once training and learning was complete across models, results were aggregated into a dictionary in python where they were assigned to their respective body part during testing (Table 1, Figure 2). In the end, the developed application assigned binary numerical values across traits for any new specimen for which images that were not used as a training set. For each trait, the most frequently occurring value (mode) across all images in the folder is calculated. Mode values are stored in a Python dictionary as one entry per trait. The final output includes the mode for each model and a sex prediction. Thus, the species identity is determined by summing the results for every model except the Sex classifier, with a minimum score of 0 and a maximum score of 9. If the final score is ≤ 3 or ≥ 7, the duck would be classified as a Mexican Duck or Mallard, respectively, with in-between scores labeled as Unknown. These thresholds were chosen to represent conservative boundaries that minimize misclassification, recognizing that individuals with intermediate scores likely reflect either hybrid ancestry or ambiguous phenotypic expression. Finally, sex was determined using the bill color model with an output of 0 or 1

indicating female or male, respectively. Overall, we evaluated classification performance at the individual level by combining predictions from the 10 trait-specific classifiers post-hoc using the consensus scoring procedure as s sum across traits and using the class with the highest cumulative score to assign the final label.

Finally, to calculate the overall confidence of the classification models, we used the raw prediction outputs (logits) generated by running each image through the trained *fast.ai* visual learner models saved as *.pkl* files. These logits were converted into class probabilities using the softmax function, which normalizes the outputs so that the sum across all classes equals one. For each prediction, we determined the predicted class by applying the argmax function that selects the index of the class with the highest probability of classification. The probability corresponding to the predicted class was then extracted and treated as a confidence score for that prediction. Because multiple predictions were made for each model, a list of confidence scores was collected for each model. From these lists, the highest confidence value per model was selected to represent the maximum certainty achieved by that model. To obtain a single measure of confidence across all models, the maximum confidence values from each model were averaged. This average of per-model maximum confidences was then expressed as a percentage (a.k.a., overall accuracy) to provide an interpretable summary of the overall model confidence.

Finally, we attempted to test model performance by having three non-experts and two experienced field biologists independently classify the same 30 images evaluated by the CNN for species, sex, and full identification (species + sex). Classification accuracies were recorded to contextualize the CNN's performance relative to expert identification.

2.1 Heatmaps and trait validation

Once models were trained, a set of nine randomly chosen images of the same body region that the trait is testing on are printed to show the prediction and true label for that image (Figure 3). Doing so provides clarity into whether the computer vision is correctly scanning the trait of interest, as well troubleshoot when scores and predictions are inaccurate (Figure 1). To do this, we generate an occlusion map using the Captum's IntegratedGradients package (Kokhlikyan et al, 2020) that was specifically developed for visualization of ResNet models. This occlusion map takes a test picture and creates a heatmap of where in the image the model focuses most for its prediction (Figure 2), and where the darker color on the image demarcated highest model focus. Finally, a confusion matrix is generated to inspect the amount of type 1 and type 2 error(s). This protocol is repeated for each individual trait, using the previously reported body region folder as the training input for each trait model.

3. Results

Each trait model was trained and validated on a unique body-specific images of genetically known Mexican Ducks or Mallards ($N_{Mallard}$ = 1268 images; $N_{Mexican\ Duck}$ = 2824 images) with an 80%-20% train-validation split (362 – 958 images for training, 74 – 239 for validation; see Supplementary Table S1 for per-body region counts). Across the 10 trait-specific classifiers, balanced accuracies during internal validation ranged from 78% to 100% (Table 2), where balanced accuracy evaluates the network's performance when dealing with imbalanced datasets, with most models performing extremely well. Models trained on high-contrast or structurally distinct traits achieved >97% balanced accuracy. For example, Speculum Color model reliably classified the presence of iridescent green or blue (balanced accuracy = 100%, see Figure 3A), and the Overall Breast and Belly plumage model effectively distinguished between solid and mottled patterns (balanced accuracy = 100%, see Figure 3B). In contrast, models trained on presence of

green on the head and Greater Secondary Covert Patterns produced lower balanced accuracies (range: 78–92%; see Table 2 and Supplementary Table S2). Finally, the sex classification model achieved a balanced accuracy of 93%, but sex was not used as part of the final score calculation for species classification.

To evaluate generalization beyond the training dataset, the complete set of trait models were validated with 30 new genetically-known individuals that were entirely unseen during training. When combined to classify species for these individuals, the integrated trait-based system achieved an overall balanced accuracy of 72% with precision, recall, and F1 scores of 50%, 67%, and 57% respectively. The model resulted in a full identification accuracy (where both sex and species were correct) of 40.00%, sex accuracy of 53.33% and species accuracy of 66.67% (Table 3). AI model accuracies were closer to non-expert people achieved 43-56% accuracy in full identification and 66.7-76.7% species accuracy as compared to experts who reached accuracies of 86.7% full identification, 90% for sex, and 86% for species (Table 4, Supplementary Tables S3-S8).

Next, we applied IntegratedGradients attribution heatmaps (Kokhlikyan et al., 2020) across trait models to understand regions of model activity. Although, heatmaps consistently highlighted biologically relevant regions in several models (e.g., sex, belly models; Figure 3A), others showed activation outside the primary trait region (e.g., the speculum model also included adjacent primary feathers; Figure 3B). For example, whereas the sex model concentrated activation on the bill and head plumage as desired (Figure 3C), background or non-biological elements (e.g., hands, tags, shoes, flooring) were detected among the top attribution regions in six (of 10) trait-specific heatmaps and despite image segmentation (Supplementary Figure S2).

4. Discussion

The application of convolutional neural networks towards species conservation is increasingly inevitable as ecological research shifts towards data-intensive, image-based monitoring (Lamba et al, 2019; Ullah et al, 2025). The exponential growth of digital imagery from community science platforms, camera traps, and museum collections provides a natural foundation for machine learning applications, while the need for rapid, scalable, and objective species identification makes deep learning approaches particularly well suited for conservation science. However, appropriate dataset size and accuracy, as well as model development, remain key limitations before their full potential can be realized. Existing tools such as Merlin & iNaturalist rely on extensive visual datasets requiring substantial expert annotation, an effort often infeasible for rare or cryptic species, and may embed unknown observer biases. Reported classification accuracies in generalist ecological CNN systems vary widely, often ranging from 70-97% for species-level identification when trained on very large and visually redundant datasets (Binta Islam et al., 2023; Norouzzadeh et al., 2018). More specialized bird-identification models trained only on external phenotypes similarly achieve high accuracy primarily when dataset sizes exceed tens or hundreds of thousands of labeled images. In comparison, the individual model balanced accuracies obtained here (>90%) using a genetically vetted dataset of only 3,301 images demonstrate that biologically grounded training labels can substantially compensate for smaller sample sizes. Although direct comparison is imperfect, given that studies do not generally incorporate underlying genetic structure, the results indicate competitive performance relative to large-scale visual-only systems while offering greater taxonomic rigor. In most applications, model optimization focuses primarily on expanding dataset size to overcome variability within smaller samples. Still, interpretability-based approaches offer an

alternative path, reducing the need for such massive datasets by identifying how and where models visualize patterns (Yosinski et al, 2015).

Here we introduce a novel and interpretable CNN framework specifically designed for ecological application. Rather than training models on visually identified individuals, we use genetically verified images as the foundation for trait recognition, ensuring that learned features reflect true biological difference rather than superficial phenotypic variability. This approach directly addresses a major limitation in previous species identification efforts that rely solely on external appearance which is particularly problematic for closely related species or hybrids where plumage traits can be ambiguous (Brown et al., 2022b). By pairing genetically partitioned datasets with trait-specific CNN classifiers, we demonstrate that high accuracy (balanced accuracies >90%; Table 2) comparable to expert-based identification can be achieved even with limited dataset of 3,301 images. When compared to human ability, although the CNN does not yet reach expert-level performance, it matches or exceeds non-expert capabilities and thus provides a promising baseline for automated identification – particularly in scenarios where skilled field experts are unavailable (Table 4).

Among improvements, the application of heatmaps proved to identify innate challenges during trait learning by the simple nature of image interpretability among convolutional neural networks (Figure 3). Models struggled to focus on specific traits even when individuals were partitioned by parts of the birds first (Figure 2), with models often marking background or foreground objects unrelated to the bird of interest. For example, we found that due to many Mexican Duck images taken with a white background surrounding the duck caused test images with white background to be identified as a Mexican Duck with high confidence regardless of subject identity. Understanding this, we was able to include steps for further image partitioning

through background segmentation and removal to circumvent this issue. Finally, although the multi-network approach undertaken to emphasize interpretability may limit performance achieved from a single, standard CNN, we argue that the tradeoff is necessary when working with relatively small to moderate datasets. Specifically, application of outlined data augmentation steps of cropping, hue and value shifts, and rotations to our modest dataset of 2,824 and 1,268 Mexican Duck and Mallard images, respectively, proved to overcome the large datasets often required for training of CNNs (Uchida et al, 2016); however, we acknowledge that while predictive accuracies were promising, further improvements is expected with larger, more diverse datasets.

Beyond qualitative interpretability, formal analysis of model error and uncertainty represents an important opportunity for future optimization. Misclassification patterns in this study, such as systematic confusion driven by background artifacts, suggest that probabilistic outputs (i.e. the softmax-derived confidence scores for each class) could be leveraged to identify image regions or trait categories that require targeted augmentation, improved segmentation, or rebalancing. Incorporating uncertainty-aware methods, such as Monte Carlo dropout (Asgharnezhad et al., 2025) or deep ensembles (Zhou et al., 2024), would enable quantification of prediction confidence and help distinguish between errors driven by insufficient data versus genuinely ambiguous phenotypes. Such approaches could guide iterative dataset expansion by highlighting which regions of phenotype space require additional genetically verified training images.

4.1 Application in Wildlife Conservation & Future Work

Wildlife conservation is increasingly challenged by accelerating biodiversity loss, habitat fragmentation, and the growing need for efficient monitoring at scale. Traditional field methods

remain invaluable but are often labor intensive, time consuming, and limited in geographic scope. Consequently, artificial intelligence is being recognized as a transformative tool in wildlife conservation for species identification (Sharma et al, 2022; Soni et al, 2023) and population monitoring (Hedge et al, 2024; Kumar & Jakhar, 2022). Generally, field key development can achieve sufficiently high accuracies, but depend on manual measurements and notes that require some expertise, and thus, remain limiting for more rapid and large-scale deployment (Brown et al, 2022b). Rather, our approach no longer requires field expertise by coupling a field key-like scoring system with computer vision that enables near-instantaneous identification of individuals from photos; though the initial scoring system that the models are trained on still require expertise to ensure high-quality training. Importantly, validation of accuracy is not limited to model scores, but the inclusion of genetic information provides a secondary validation step. In short, building models based on genetically vetted individuals not only ensures that any nuance of phenotypic variability due to hybridization does not further complicate models up front, but also will eventually permit for a scoring system to output genetic probabilities. However, hybrid classification remains outside the present scope of this study, as current trait-based classifiers are not yet sufficiently robust to reliably identify mixed and intermediate phenotypes.

We acknowledge that further training of our models with additional individuals will be required to increase accuracy scores, as well as training models with longer epochs on dedicated GPU hardware (which is currently out of our budget's scope) will further improve classification accuracy. Nevertheless, our approach can be extrapolated to any species that uses trait-specific classifiers and a moderate, but high-quality dataset of images to train with that can facilitate real-time species differentiation in the field. Moreover, developing a system grounded in genetically confirmed data would elevate the reliability of training datasets, ensuring that ground-truth labels

are accurate rather than reliant on potentially ambiguous visual identification. In addition, establishing a high-quality reference dataset of individuals with known genetic lineages would allow for true categorical classification of parental taxa, providing stronger alignment between genetic and image-based datasets. Next, integrating advanced image segmentation approaches (i.e., Mask R-CNN; He et al, 2017) with a ResNet backbone can help isolate relevant bird features prior to classification. Doing so would minimize the influence of background noise and sharpen model attention on diagnostically important traits. Model architecture also presents opportunities for performance gains. Originally developed for natural language processing, transformer-based models are now used in computer vision operate by modeling long-range relationships across an image through self-attention mechanisms (Khan et al., 2022), enabling them to capture subtle, spatially distributed features that standard CNNs may overlook. Because their architecture excels at fine-grained pattern recognition, models such as Vision Transformers (ViT) (Dosovitskiy et al., 2021) and Swin Transformers (Liu et al., 2021), alongside advanced CNN architectures like EfficientNet (Tan & Le, 2020), may be particularly well suited for the trait-specific and fine-scale discrimination needed for species identification in similar-looking taxa. Their ability to extract multi-scale features and integrate information across the entire image suggests strong potential for improving classification of intermediate, ambiguous or hybrid phenotypes. Similarly, lightweight architectures such as MobileNetV3 (Howard et al., 2019) and ConvNeXt (Liu et al., 2022) could support deployment in field applications with limited hardware while maintaining high accuracy. Future work should evaluate these architectures using the genetically confirmed dataset established here to determine whether more expressive backbones can further close the performance gap with expert observers. Finally, the potential of extracting numerical values for traits for individuals with

362 genetic information allows the possibility of understanding complex trait evolution and expression
363 through genome-wide association studies (Sella & Barton, 2019).

Literature Cited

Adolph, S. C., & Hardin, J. S. (2007). Estimating phenotypic correlations: correcting for bias due to intraindividual variability. *Functional Ecology*, 178-184.

Asgharnezhad, H., Shamsi, A., Alizadehsani, R., Mohammadi, A., & Alinejad-Rokny, H. (2025). *Enhancing Monte Carlo Dropout Performance for Uncertainty Quantification* (arXiv:2505.15671). arXiv. https://doi.org/10.48550/arXiv.2505.15671

Atanbori, J., Duan, W., Murray, J., Appiah, K., and Dickinson, P (2016). "Automatic classification of flying bird species using computer vision techniques," *Pattern Recognit. Lett*., vol. 81, pp. 53–62, doi: 10.1016/j.patrec.2015.08.015.

Binta Islam, S., Valles, D., Hibbitts, T. J., Ryberg, W. A., Walkup, D. K., & Forstner, M. R. J. (2023). Animal Species Recognition with Deep Convolutional Neural Networks from Ecological Camera Trap Images. *Animals*, *13*(9), 1526. https://doi.org/10.3390/ani13091526

Brandes, T (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1), S163-S173.

Brenowitz, E. A., and Larson, T. A. "Neurogenesis in the adult avian song-control system." *Cold Spring Harbor perspectives in biology* vol. 7,6 a019000. 1 Jun. 2015, doi:10.1101/cshperspect.a019000

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The balanced accuracy and its posterior distribution. In 2010 20th international conference on pattern recognition (pp. 3121-3124). *IEEE*.

Brown, J. I., Harrigan, R. J., & Lavretsky, P. (2022). Evolutionary and ecological drivers of local adaptation and speciation in a North American avian species complex. *Molecular Ecology*, *31*(9), 2578-2593.

Brown, J. I., Hernández, F., Engilis Jr, A., Hernández-Baños, B. E., Collins, D., & Lavretsky, P (2022). Genomic and morphological data shed light on the complexities of shared ancestry between closely related duck species. *Scientific reports*, *12*(1), 10212.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A (2020). Albumentations: fast and flexible image augmentations. *Information*, *11*(2), 125.

Chu, M (2012). Merlin: Online bird identification with human learning and machine learning.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Duong, D., Waikel, R. L., Hu, P., Tekendo-Ngongang, C., & Solomon, B. D (2022). Neural network classifiers for images of genetic conditions with cutaneous manifestations. *Human Genetics and Genomics Advances*, *3*(1).

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, 68(8), 563-576.

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*, 107268.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the *IEEE International Conference on Computer Vision* (pp. 2961-2969).

He, K., Zhang, X., Ren, S., & Sun, K (2015). Deep Residual Learning for Image Recognition. *arXiv* preprint arXiv:1512.03385.

Hegde, N. R., & Bargavi, D. S. (2024). Smart Conservation: Integrating AI for enhanced wildlife monitoring. *International Research Journal of Computer Science*.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). *Searching for MobileNetV3* (arXiv:1905.02244)

Howard, J., & Gugger, S. (2020). Fastai: a layered API for deep learning. *Information*, 11(2), 108.

Incze, A., Jancsó, H. B., Szilágyi, Z., Farkas, A., & Sulyok, C. (2018). Bird sound recognition using a convolutional neural network. In 2018 IEEE 16th international symposium on intelligent systems and informatics (SISY) (pp. 000295-000300). IEEE.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10s), 1–41. https://doi.org/10.1145/3505244

Kirby, R. E., Reed, A., Dupuis, P., Obrecht, H. H. III, & Quist, W. J. (2000). Description and identification of American black duck, mallard, and hybrid wing plumage (p. 26). *Biological Science Report. U.S. Geological Survey, Biological Resources Division*.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for PyTorch* (arXiv:2009.07896).

433 Kumar, D., & Jakhar, S. D. (2022). Artificial intelligence in animal surveillance and
434 conservation. *Impact of artificial intelligence on organizational transformation*, 73-85.
435 Lamba, A., Cassey, P., Segaran, R. R., & Koh, L. P. (2019). Deep learning for environmental
436 conservation. *Current Biology*, 29(19), R977-R982.
437 Lavretsky, P (2021). Population genomics provides key insights into admixture, speciation, and
438 evolution of closely related ducks of the mallard complex. *Population Genomics:*
439 *Wildlife*, 295-330.
440 Lavretsky, P., DaCosta, J. M., Hernández-Baños, B. E., Engilis, A., Sorenson, M. D., & Peters, J.
441 L. (2015). Speciation genomics and a role for the Z chromosome in the early stages of
442 divergence between Mexican ducks and mallards. *Molecular Ecology*, *24*(21), 5364–
443 5378. https://doi.org/10.1111/mec.13402
444 Lavretsky, P., DaCosta, J. M., Sorenson, M. D., McCracken, K. G., & Peters, J. L (2019).
445 ddRAD-seq data reveal significant genome-wide population structure and divergent
446 genomic regions that distinguish the mallard and close relatives in North
447 America. *Molecular Ecology*, *28*(10), 2594-2609.
448 Lavretsky, P., Duenez, E., Kneece, M., & Kaminski, R. M (2021). Population genetics of a
449 translocated population of Mottled Ducks and allies. *The Journal of Wildlife*
450 *Management*, *85*(8), 1616-1627.
451 LeCun, Y., Bengio, Y., & Hinton, G (2015). Deep learning. *Nature*, *521*(7553), 436–444.
452 https://doi.org/10.1038/nature14539
453 Lipton, Z. C (2018). The mythos of model interpretability: In machine learning, the concept of
454 interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

Liu, L., & Du, K. (2024). A perspective on computer vision in biosensing. *Biomicrofluidics*, *18*(1), 011301. https://doi.org/10.1063/5.0185732

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows* (arXiv:2103.14030).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the 2020s* (arXiv:2201.03545).

Lucas, T. C (2020). A translucent box: interpretable machine learning in ecology. *Ecological Monographs*, *90*(4), e01422.

Man, K., & Chahl, J (2022). A review of synthetic image data and its use in computer vision. *Journal of Imaging*, *8*(11), 310.

Marini, A., Facon, J., and Koerich, A. L (2013). "Bird Species Classification Based on Color Features," *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 4336–4341, doi: 10.1109/SMC.2013.740.

McCartney, K. R., Kumar, S., Sing, S. E., & Ward, S. M (2019). Using invaded-range species distribution modeling to estimate the potential distribution of Linaria species and their hybrids in the US northern Rockies. *Invasive Plant Science and Management*, *12*(2), 97-111.

Mehyadin, A. E., Abdulazeez, A. M., Hasan, D. A., & Saeed, J. N (2021). Birds sound classification based on machine learning algorithms. Asian Journal of Research in Computer Science, 9(4), 1-11.

476 Mendes, E.D., Wooley, J., Pi, Y., Tao, J., & Tedeschi, L.O (2024). 419 Automated individual
477 animal identification and feeding bunk scoring: a computer vision approach for beef
478 cattle at Calan gate feeding system. *Journal of Animal Science*.
479 Miao, Z., Gaynor, K. M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M. S., McInturff, A.,
480 Bowie, R. C. K., Nathan, R., Yu, S. X., & Getz, W. M. (2019). Insights and approaches
481 using deep learning to classify wildlife. *Scientific Reports*, *9*(1), 8137.
482 Moallem, G. (2021). Applications of image processing and machine learning techniques in
483 wildlife monitoring and cancer cell characterization (Doctoral dissertation).
484 Müller, T., Philippi, N., Dandekar, T., Schultz, J., & Wolf, M. (2007). Distinguishing
485 species. *Rna*, *13*(9), 1469-1472.
486 Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine
487 learning: definitions, methods, and applications. *arXiv preprint arXiv*:1901.04592.
488 Nadimpalli, U. D., (2005). "Image processing techniques to identify predatory birds in
489 aquacultural settings" LSU Master's Theses. 276.
490 Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., &
491 Clune, J. (2018). Automatically identifying, counting, and describing wild animals in
492 camera-trap images with deep learning. *Proceedings of the National Academy of*
493 *Sciences*, *115*(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115
494 Perry, W. L., Lodge, D. M., & Feder, J. L (2002). Importance of hybridization between
495 indigenous and nonindigenous freshwater species: an overlooked threat to North
496 American biodiversity. *Systematic Biology*, *51*(2), 255-275.
497 Picard, S., Chapdelaine, C., Cappi, C., Gardes, L., Jenn, E., Lefèvre, B., & Soumarmon, T
498 (2020). Ensuring dataset quality for machine learning certification. In *2020 IEEE*

international symposium on software reliability engineering workshops (ISSREW). (pp. 275-282). IEEE.

Reyes, D., & Sánchez, J (2024). Performance of convolutional neural networks for the classification of brain tumors using magnetic resonance imaging. *Heliyon*, *10*(3).

Sella, G., & Barton, N. H. (2019). Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annual review of genomics and human genetics*, 20(1), 461-493.

Sharma, S., Sato, K., & Gautam, B. P. (2022). Bioacoustics Monitoring of Wildlife using Artificial Intelligence: A Methodological Literature Review. *2022 International Conference on Networking and Network Applications (NaNA)*, 1–9. https://doi.org/10.1109/NaNA56854.2022.00063

Smith, L. N., & Topin, N (2019, May). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications* (Vol. 11006, pp. 369-386). SPIE.

Soni, P., Dhavale, S., Yenishetti, S., Panat, L., & Karajkhede, G. (2023, October). Medicinal plant species identification using AI. *In 2023 IEEE 11th Region 10 Humanitarian Technology Conference* (R10-HTC) (pp. 662-668). IEEE.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol Skrifter/Kongelige Danske Videnskabernes Selskab*., 5, 1.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

Tan, M., & Le, Q. V. (2020). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* (arXiv:1905.11946). arXiv. https://doi.org/10.48550/arXiv.1905.11946

Uchida S., Ide, S., Iwana, B. K., and Zhu, A (2016). A Further Step to Perfect Accuracy by Training CNN with Larger Data. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR*). 405-410, doi: 10.1109/ICFHR.2016.0082.

Ueda, K., & Loarie, S (2008). iNaturalist. Retrieved from https://www.inaturalist.org/

Ullah, F., Saqib, S., & Xiong, Y. C. (2025). Integrating artificial intelligence in biodiversity conservation: bridging classical and modern approaches. *Biodiversity and Conservation*, 34(1), 45-65.

Van Der Sluis, S., Verhage, M., Posthuma, D., & Dolan, C. V. (2010). Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PloS one*, *5*(11), e13929.

Wells, C. P., Lavretsky, P., Sorenson, M. D., Peters, J. L., DaCosta, J. M., Turnbull, S., ... & Engilis Jr, A (2019). Persistence of an endangered native duck, feral mallards, and multiple hybrid swarms across the main Hawaiian Islands. *Molecular Ecology*, *28*(24), 5203-5216.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding Neural Networks Through Deep Visualization (No. arXiv:1506.06579). *arXiv*. https://doi.org/10.48550/arXiv.1506.06579

Yoss, A. (2020). Transfer learning using pre-trained AlexNet model and fashion-MNIST. *Towards Data Science*. https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96

Zhou, Y., Zhu, H., Chai, Y., Jiang, Y., & Liu, Y. (2024). *Towards Trustworthy Web Attack Detection: An Uncertainty-Aware Ensemble Deep Kernel Learning Model* (arXiv:2410.07725). arXiv. https://doi.org/10.48550/arXiv.2410.07725

Table 4. Accuracy of identification of the same 30 individuals, compared by AI models, non-experts and experts.

	AI CORRECT	AI INCORRECT	AI % ACCURACY
FULL ID	12	18	40.00%
SEX	16	14	53.33%
SPECIES	20	10	66.67%

	NON EXPERT 1 CORRECT	NON EXPERT 1 INCORRECT	NON EXPERT 1 % ACCURACY	NON EXPERT 2 CORRECT	NON EXPERT 2 INCORRECT	NON EXPERT 2 % ACCURACY	NON EXPERT 3 CORRECT	NON EXPERT 3 INCORRECT	NON EXPERT 3 % ACCURACY
FULL ID	13	17	43.33%	15	15	50.00%	17	13	56.67%
SEX	20	10	66.67%	20	10	66.67%	20	10	66.67%
SPECIES	20	10	66.67%	23	7	76.67%	26	4	86.67%

	EXPERT 1 CORRECT	EXPERT 1 INCORRECT	EXPERT 1 % ACCURACY	EXPERT 2 CORRECT	EXPERT 2 INCORRECT	EXPERT 2 % ACCURACY
FULL ID	26	4	86.67%	25	5	83.33%
SEX	27	3	90.00%	30	0	100.00%
SPECIES	26	4	86.67%	24	6	80.00%

Table 3. Aggregate AI model classification results for species and sex identification, along with the according score and final model confidence. Text in red highlights an incorrect identification.

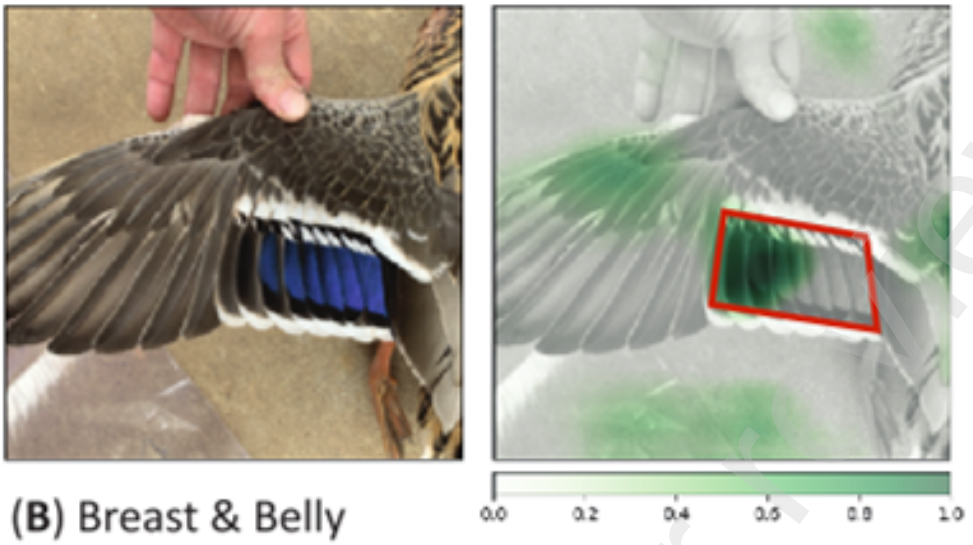
Sample	Correct Species	Identified Species	Correct Sex	Identified Sex	Score	Confidence
1	Mallard	Mexican Duck	Female	Female	4	77%
2	Mallard	Mexican Duck	Male	Female	3	70%
3	Mallard	Mallard	Male	Female	8	71%
4	Mexican Duck	Mexican Duck	Male	Male	4	79%
5	Mexican Duck	Mexican Duck	Male	Male	3	82%
6	Mexican Duck	Mexican Duck	Female	Male	4	87%
7	Mallard	Mallard	Female	Female	7	75%
8	Mallard	Mallard	Male	Male	8	71%
9	Mexican Duck	Mexican Duck	Male	Female	3	70%
10	Mexican Duck	Mallard	Male	Female	9	78%
11	Mexican Duck	Unknown	Male	Male	6	81%
12	Mexican Duck	Mexican Duck	Female	Male	1	75%
13	Mexican Duck	Mexican Duck	Female	Female	2	81%
14	Mallard	Mallard	Male	Female	3	74%
15	Mallard	Unknown	Male	Female	3	80%
16	Mallard	Mallard	Male	Male	3	71%
17	Mallard	Mallard	Male	Male	3	75%
18	Mallard	Unknown	Female	Male	0	83%
19	Mallard	Unknown	Female	Male	0	77%
20	Mallard	Mallard	Male	Male	3	77%
21	Mallard	Mexican Duck	Female	Male	0	77%
22	Mallard	Mallard	Male	Male	3	79%
23	Mallard	Mallard	Female	Male	3	79%
24	Mexican Duck	Mexican Duck	Male	Male	0	76%
25	Mexican Duck	Mexican Duck	Male	Male	0	72%
26	Mexican Duck	Unknown	Female	Female	3	78%
27	Mexican Duck	Unknown	Male	Male	0	77%
28	Mexican Duck	Mexican Duck	Male	Female	0	79%
29	Mexican Duck	Mexican Duck	Female	Male	0	72%
30	Mexican Duck	Mexican Duck	Male	Male	0	71%

Full Species		
Correct	12	40.00%
Incorrect	18	60.00%
Sex		
Sex Correct	16	53.33%
Sex Incorrect	14	46.67%
Species		
Species Correct	20	66.67%
Species Incorrect	10	33.33%

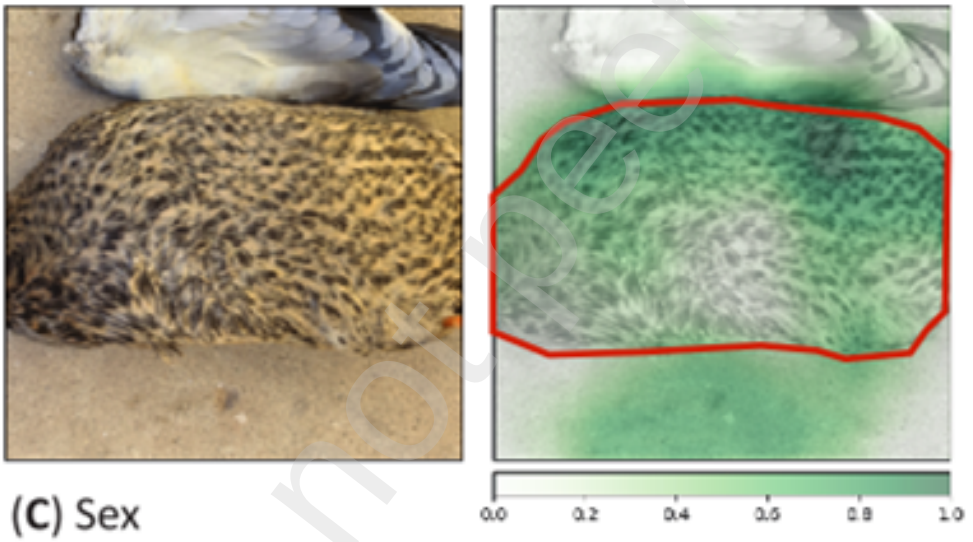
Figure 2. Graphical representation of the body regions used for training individual trait-specific models. *Art credit to August Konvalin.*

Figure 3. Example of heatmap integration within the (A) speculum, (B) belly, and (C) sex models by using the IntegratedGradients package, demonstrating interpretability through visualization of regional activation of each model. Note that while traits of interest were focused on, activation of backgrounds and non-trait regions were evident across models

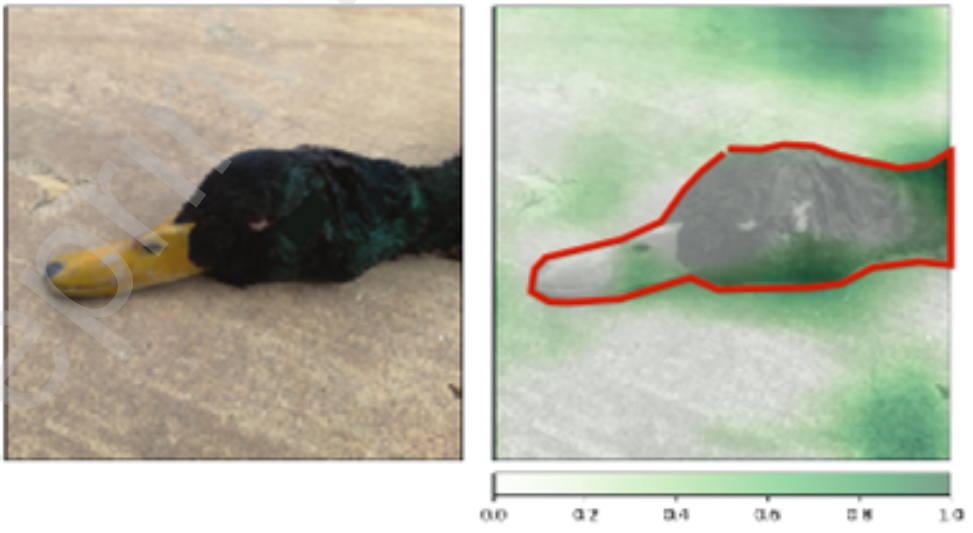
(A) Speculum

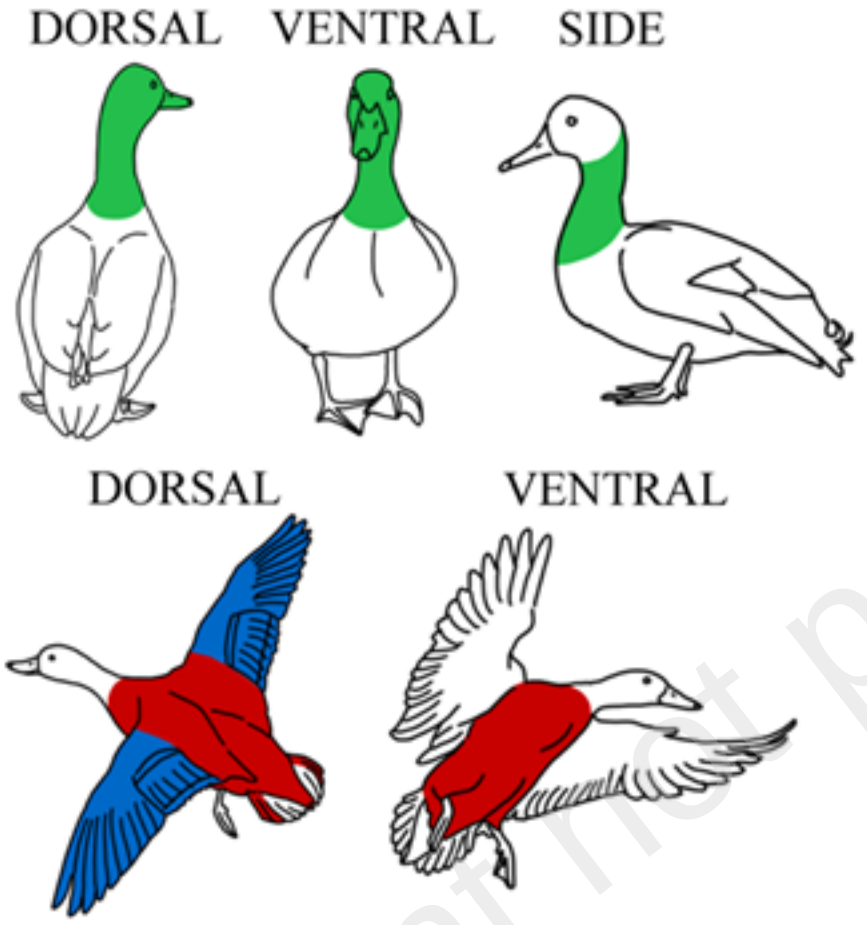


(B) Breast & Belly



(C) Sex





HEAD
= % GREEN
= FACE/NECK
= SEX

WING
= PRIMARY COVERT
= LESSER COVERT
= GREATER SECONDARY
= SPECULUM COLOR

BODY
= BACK PATTERN
= OUTER TAIL FEATHERS
= BREAST/BELLY

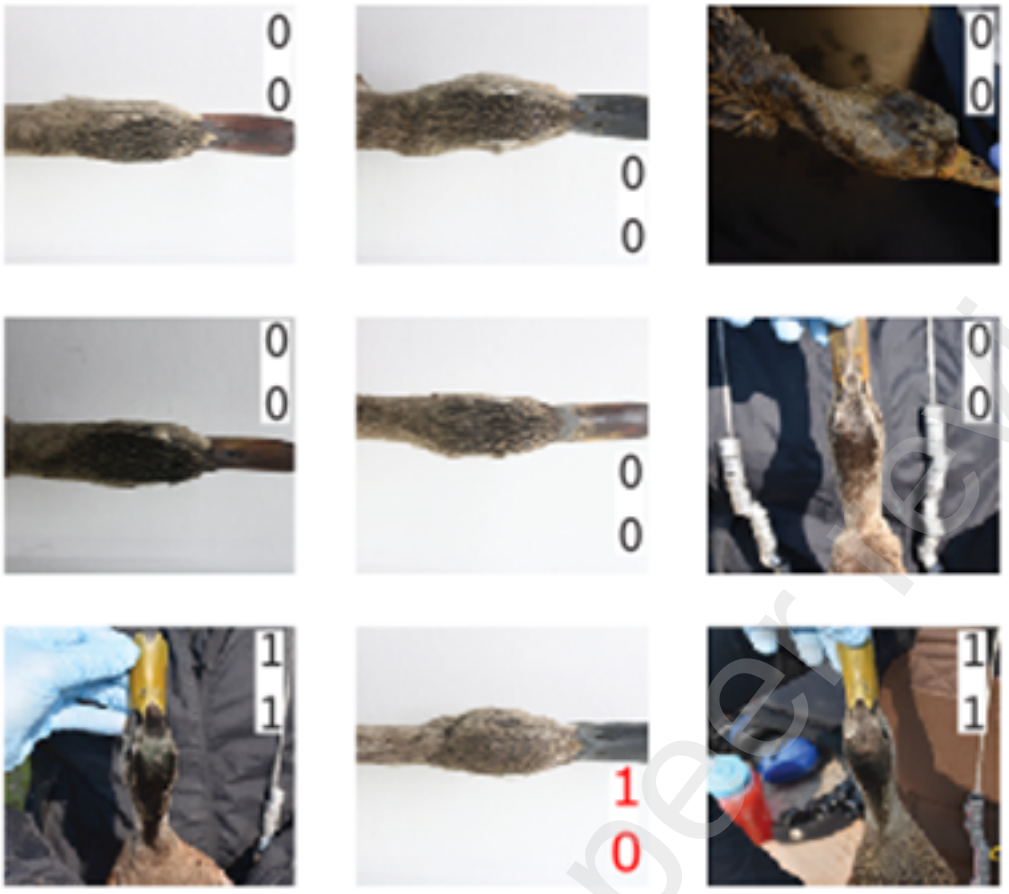


Figure 1. Per trait-specific model Confusion matrices for the models, here showing the matrix for the sex classifier. 9 images were selected from the test pool with the model's predictions as well as the correct label printed under the prediction. In this case, out of the 9 images, only the bottom middle was incorrectly labelled as male (1) when it should have been female (0).