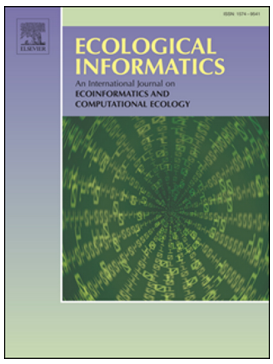
Discrimination between the facial gestures of vocalising and non-vocalising lemurs and small apes using deep learning



Discrimination between the facial gestures of vocalising and non-vocalising lemurs and small apes using deep learning

Filippo Carugati, Olivier Friard, Elisa Protopapa, Camilla Mancassola, Emanuela Rabajoli, Chiara De Gregorio, Daria Valente, Valeria Ferrario, Walter Cristiano, Teresa Raimondi, Valeria Torti, Brice Lefaux, Longondraza Miaretsoa, Cristina Giacoma, Marco Gamba

PII: S1574-9541(24)00389-3

DOI: https://doi.org/10.1016/j.ecoinf.2024.102847

Reference: ECOINF 102847

To appear in: *Ecological Informatics*

Received date: 11 April 2024

Revised date: 3 October 2024

Accepted date: 4 October 2024

Please cite this article as: F. Carugati, O. Friard, E. Protopapa, et al., Discrimination between the facial gestures of vocalising and non-vocalising lemurs and small apes using deep learning, *Ecological Informatics* (2024), https://doi.org/10.1016/j.ecoinf.2024.102847

DISCRIMINATION BETWEEN THE FACIAL GESTURES OF VOCALISING AND NON-VOCALISING LEMURS AND SMALL APES USING DEEP LEARNING

Filippo Carugati[1, *], Olivier Friard[1], Elisa Protopapa[1], Camilla Mancassola[1], Emanuela Rabajoli[1], Chiara De Gregorio[1], Daria Valente[1,2], Valeria Ferrario[1,3], Walter Cristiano[1], Teresa Raimondi[1], Valeria Torti[1], Brice Lefaux[4], Longondraza Miaretsoa[5,6], Cristina Giacoma[1], Marco Gamba[1, *]

1 Department of Life Sciences and Systems Biology, Università di Torino, Torino, Italy.

2 Parco Natura Viva Garda Zoological Park, 37012 Bussolengo, Italy

3 Chester Zoo, Caughall Road, Chester, UK

4 Zoo de Mulhouse, Mulhouse, France.

5 Groupe d'Etude et de la Recherche sur les Primates de Madagascar (GERP)

6 Institut Supérieur de Technologie Régionale (ISTR), Manakara, Madagascar

*Corresponding authors:

Filippo Carugati (filippo.carugati@unito.it)

Via Accademia Albertina 13, 10123, Torino, Italy

Marco Gamba (marco.gamba@unito.it)

ORCID:

Filippo Carugati: 0000-0002-5754-5787

Olivier Friard: 0000-0002-0374-9872

Elisa Protopapa: 0009-0004-9032-8428

Camilla Mancassola: 0009-0009-6134-5911

Emanuela Rabajoli: 0009-0008-2498-0601

Chiara De Gregorio: 0000-0001-7017-6181

Daria Valente: 0000-0001-6086-5135

Valeria Ferrario: 0000-0002-7958-738X

Walter Cristiano: 0000-0002-2634-9716

Teresa Raimondi: 0000-0001-6767-1835

Valeria Torti: 0000-0002-6908-1203

Longondraza Miaretsoa: 0000-0002-0403-0982

Cristina Giacoma: 0000-0002-8429-7723

Marco Gamba: 0000-0001-9545-2242

*Corresponding authors:

Filippo Carugati

Department of Life Sciences and Systems Biology, Università di Torino, Torino, Italy.

Mail: filippo.carugati@unito.it

Marco Gamba

Department of Life Sciences and Systems Biology, Università di Torino, Torino, Italy.

Mail: marco.gamba@unito.it

Data Availability.

The data and the code are available at https://zenodo.org/records/13853817

Competing interests.

We declare we have no competing interests.

Acknowledgements.

We are grateful to GERP (Groupe d'Étude et des Recherche sur les Primates de Madagascar) for their help in organising the field activities. We thank Dr. Cesare Avesani Zaborra and Dr. Caterina Spiezio for their support. We thank the local field guides for their help and logistical support during the data collection. We thank Dr. Stefano Zucca and Prof. Serena Bovetti for letting us use the computer "SuperBrain2" to train the models and data analysis. We also thank Sergio Rabellino and the C3S staff for supporting us using the OCCAM computational facilities. Finally, we thank the anonymous reviewers whose helpful comments and suggestions critically improved our work.

Author contribution.

Filippo Carugati: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review and Editing, Visualization. **Olivier Friard**: Methodology, Software, Validation, Data Curation, Writing – Review and Editing. **Elisa Protopapa**: Methodology, Investigation, Writing – Review and Editing. **Camilla Mancassola**: Methodology, Investigation, Writing – Review and Editing. **Emanuela Rabajoli**: Methodology, Investigation, Writing – Review and Editing. **Chiara De Gregorio:** Writing – Review and Editing. **Daria Valente**: Writing – Review and Editing. **Valeria Ferrario**: Writing – Review and Editing. **Walter Cristiano**: Writing – Review and Editing. **Teresa Raimondi**: Writing – Review and Editing. **Valeria Torti**: Writing – Review and Editing. **Brice Lefaux**: Resources, Writing – Review and Editing. **Longondraza Miaretsoa**: Resources, Writing – Review and Editing. **Cristina Giacoma**: Resources, Writing – Review and

Editing. **Marco Gamba**: Conceptualization, Methodology, Software, Formal Analysis, Supervision, Writing – Original draft, Writing – Review and Editing

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ABSTRACT

Facial expression studies in animal communication are essential. However, manual inspection methods are only practical for small datasets. Deep learning techniques can help discriminate facial configurations associated with vocalisations over large datasets. We extracted and labelled frames of different primate species, trained deep-learning models to identify key points on their faces, and computed distances between them to identify facial gestures. We used machine learning algorithms to classify vocalised and non-vocalised gestures across different species. The algorithms showed higher-than-chance correct classification rates, with some exceeding 90%. Our work employs deep learning to map primate facial gestures and offers an innovative application of pose estimation systems. Our approach facilitates the investigation of facial repertoire across primate species and behavioural contexts, enabling comparative research in primate communication.

Keywords: primate face, *Indri indri, Propithecus diadema, Nomascus gabriellae,* DeepLabCut, acoustic communication

1. INTRODUCTION

Advances in recording instruments have pushed the study of animal behaviour towards applying technologies that enable the processing of large volumes of data (Steenweg et al. 2017; Sugai et al. 2018; Janisch et al. 2021). The application of modern technologies is critical to reducing the mismatch between the volume of raw materials collected in the wild and captive settings and the capacity to extract meaningful information rapidly (Tuia et al., 2022). Time constraints for data analysis have prompted scholars to pay more attention to new computer technologies that can assist in analysing large amounts of video or audio recordings (Gamba et al., 2015; Friard and Gamba, 2016).

Communication studies have always been among the major themes in the study of animal behaviour. However, the study of vocal and visual communication has played a critical role in advancing our understanding of nonhuman species in the last decades (Waller et al. 2008; Parr et al. 2007). The modern approach to studying facial expression is based on the development of the Facial Action Coding System (FACS - Ekman & Friesen, 1978; Waller et al., 2020), a robust framework which allows quantifying and describing facial movements in a set of species. However, we must consider three limitations of FACS. First, identifying the FACS of a particular species requires a thorough knowledge of the anatomy underlying facial expressions (Kaminski et al., 2019). The number of species for which FACS (Facial Action Coding System) is available is limited (Waller et al., 2020). While FACS allows for detailed identification of a species' facial expressions, it requires specialised knowledge and training to recognise specific action units (Vick & Parr, 2007). Moreover, identifying each facial movement requires a manual frame-by-frame

inspection by an operator, leading to a time-consuming process that can be prone to human error (Cohn et al., 1999; Waller et al., 2022; Hamm et al., 2011).

Recent publications show that efforts are increasing to foster the development of systems that enable automatic or semi-automatic detection of facial expressions (Whitam, 2018; Morozov et al., 2021; Feilghelstein et al., 2022). This is a field of investigation in which the use of artificial intelligence (hereafter, AI) techniques is promising, along the lines of what has already been done using deep learning to extract particular acoustic signals from passive acoustic recordings (Dufourq et al. 2021; Ravaglia et al. 2023) or sequences of visual signals (e.g., the tracking of courtship flights in fruit flies - Ning et al., 2022). A new wave of studies has employed AI for facial recognition in various primate species (Guo et al., 2020), including chimpanzees (Schofield et al. 2019; Schofield et al. 2023), macaques (Paulet et al. 2024), and lemurs (Crouse et al. 2017; Deb et al. 2018). On the one hand, current approaches to using AI in developing applications that can recognise variations in facial expressions aim to automate pain recognition (rodents, Tuttle et al., 2018; sheep, Mahmoud et al., 2018; horses, Lencioni et al., 2021; cats, Feilghelstein et al. 2022). On the other hand, scientists directed efforts at the discrimination of Action Units (hereafter, AUs) to open new perspectives for social and affective neuroscience (MaqFACS, Morozov et al. 2021). Notably, AU discrimination was built on frontal head-fixed captive macaque images, thus making the replication of this pipeline challenging when applied to free-moving animals.

On a different level, markerless pose estimation dramatically progressed to capture motions (Bala et al., 2020; Hardin and Schlupp, 2022) over an increasingly large number of species thanks to the open-source software DeepLabCut (hereafter DLC, Mathis et al., 2018; Nath et al. 2019). Scholars

tracked motion and poses in various lab environments and extracted hints into the behaviour associated with particular body configurations for a diverse range of animals such as crickets (Hayakawa et al., 2024), crayfishes (Suryanto et al., 2024), dolphins (Tseng et al., 2024), rats (Popik et al., 2024; Lapp et al., 2024) and also primates (Fuchs et al., 2023). Wiltshire and colleagues (2023) also developed a robust DLC model to identify a custom set of key points distributed on the bodies of wild chimpanzees and bonobos. Perhaps because it is challenging to obtain good-quality video of the faces of free-moving animals or because it is complicated to imagine extracting facial configuration information in the presence of sometimes limited contrast between facial parts and an ever-changing background, markerless pose estimation has never been dedicated to understanding whether facial configurations are trackable.

This work aims to test the possibility of using markerless pose estimation approaches to identify given points of primate facial expressions using footage from wild and captive, freely moving animals. More precisely, we asked whether it is possible to discriminate between voiced and unvoiced facial configurations. We predict that, across primates, distinctive facial gestures were associated with phonation. Chimpanzees can emit up to 48 multimodal signals (Wilke et al. 2017), and vocal emissions are regularly accompanied by gestures and facial expressions (Tagliatela et al. 2015). Macaques associate vocalisations with particular facial postures (Hauser et al. 1993), which, in turn, are associated with vocal tract shapes (Fitch 1997).

Recognising vocalised or non-vocalised facial gestures is a pivotal step in several studies dedicated to animal communication. For instance, screening large sets of videos and pictures would allow researchers to expand our understanding of the multimodal nature of primate communication.

Also, targeting voiced gestures may enhance our knowledge of phonation mechanics and empower subtle comparative studies devoted to discovering differences within and between species or investigating particular traits related to the evolution of communication and language. Our results have important implications for using deep learning for comparative studies, starting with videos of freely moving animals.

2. MATERIAL & METHODS

2.1 Data collection

We filmed the faces of 48 individuals from three primate species, two lemur species recorded in the wild (indris - *Indri indri* - and diademed sifakas - *Propithecus diadema*) and a species of lesser ape in captivity (yellow-cheeked crested gibbons - *Nomascus gabriellae*). We collected videos of *Indri indri* and *Propithecus diadema* in the Maromizaha rainforest, Madagascar (18° 56' 49'' S, 48° 27' 53'' E) from April 26th to August 5th, 2022. We sampled ten groups of indris (for a total of 25 individuals) and five groups of diademed sifakas (for a total of 30 individuals). We followed one group per day, approximately from 6:00 AM to 1:00 PM. We recorded three yellow-cheeked crested gibbons hosted at the Zoological and Botanical Park of Mulhouse (France) between April 11th and June 17th, 2022. We filmed the individuals for 8 hours daily, from 8:30 AM to 4:30 PM. We collected all videos from outside the enclosure by placing the camera in contact with the separating glass. We recorded all videos *ad libitum* using a Panasonic Lumix FZ82 camera, equipped with a 60x zoom, that allowed us to film the subjects' faces efficiently. For all species,

we conducted recordings using an opportunist approach, filming faces whenever visible and at a distance from the operator that ranged between 2 and 20 metres.

2.2 Data preparation and training of the deep learning models

We used BORIS (Friard & Gamba, 2016) to visually inspect each video and extract clips. We selected clips using the following criteria: a minimum duration equal to 5 s, the face of a single animal was present in the shot, and no objects were standing between the operator and the animal we were filming (e.g., branches, trunks, foliage). For each clip we indicated whether the subject showed a facial configuration concomitant with a vocal emission ("*co-occurrence*", "CO") or a configuration while silent (clips labelled as *facial*, "FA"). We report the number of clips and mean duration in Table 1. We used the *FFMPEG framework* (Tomar, 2006) to convert and batch-resize the videos to a resolution of 960x540 pixels and a frame rate of 25 fps.

Species	Total clips	FA clips	CO clips	Duration (mean ± sd)
I. indri	214	162	51	14.25 ± 13.03 s
P. diadema	636	566	70	7.07 ± 4.60 s
N. gabriellae	543	429	114	44.28 ± 31.43 s

Table 1. The number of clips (total, FA and CO) extracted using BORIS and the mean (± standard deviation) duration.

We loaded the clips on DLC. We created three models, one for each target species. We used the DLC function *extract_frames* to random sample ten frames from each clip to create our training sets. Through the DLC graphical interface, we manually labelled 2355 frames for *I. indri*, 5200 for *P. diadema,* and 2370 for *N. gabriellae,* to indicate the position of a set of 13 points designed

to mark key areas of primate oro-facial configuration (see Fig. 1). We selected these points from the *primate_face* model (Witham, 2018). After a quick training, an operator could quickly and unequivocally identify these 13 markers, which can apply to a wide range of primate species. We used the coordinates of the labelled frames to train (95% of the dataset) and test (5% of the dataset) each DLC model. We used the intra-class correlation coefficient (ICC- Shrout & Fleiss, 1979) to test operators' agreement on 200 randomly selected frames extracted from 20 videos. We used two operators for each species. For all the tests, we found a high agreement between our labellers (*I. indri*: 0.986 < ICC < 0.988; *P. diadema*: 0.992 < ICC <0.993; *Nomascus* spp.: 0.997 < ICC < 0.997).

To run our deep learning model, we used a ResNet-50-based convolutional neural network (Insafutdinov et al., 2016; He et al., 2016) with default parameters for 1300000 iterations. We applied a 0.6 p-cutoff that specifies the threshold of the correct positioning likelihood. Given the GPU requirements, we trained DLC models on a computer Intel® Xeon® W-2295, 18 *core* 36 *thread*; RAM: 256 Gb; HD 12 Tb; GPU: 2x Nvidia Quadro RTX 8000 48 Gb (house name "Superbrain 2"). We ran two shuffles for each model and selected the models with the lower test error for further analysis. Once we trained the models, we used the DLC function *analyze_videos* to extract the coordinates of each key point available for every frame of all the clips.

To thoroughly test the adaptability of DLC models, we meticulously selected additional video clips (20 for *N. gabriellae*, 23 for *I. indri*, and 24 for *P. diadema*) that were not part of the initial frame extraction phase. These clips, not included in the training or test sets, could feature new animals (for *I. indri* and *P. diadema*; although, for *N. gabriellae*, it was not feasible due to fewer sampled animals) and were recorded under varying conditions such as lighting, camera angles,

environment, animal visibility, and camera distance. We randomly selected and labelled ten frames from each clip using the DLC graphical interface. Subsequently, we applied the developed models to analyse the novel videos and extract the labelled frames' coordinates. To evaluate the models' performance on novel videos, we computed the Mean Euclidean Absolute Distance (MEAD) between the coordinates of the manually labelled and predicted landmarks only when the predicted points showed a likelihood higher than the p-cutoff (0.6).

2.3 Data normalisation and preprocessing steps

Because mapping a particular frame could result in an incomplete set of markers (N < 13), we selected only the mappings featuring all the key points. Consequently, we excluded frames where at least one landmark was predicted with a likelihood lower than the p-cutoff. This situation could arise if key points were incorrectly positioned or if only parts of the face were visible, for example, when the camera angle did not allow for a direct frontal view of the subject. In this process, we also helped to minimise variability resulting from camera angle by excluding frames where excessive head rotation made it difficult to see all the points. As a result, we obtained 4240 frames for *Indri indri*, 4287 for *Propithecus diadema*, and 104111 for *Nomascus gabriellae*.

Before conducting further analysis, we performed some preprocessing steps. We could perform a facial alignment transformation or not (Feighelstein et al., 2022; Morozov et al., 2021). Facial alignment reduces the geometric variation of faces through affine transformations (e.g. point rotation) and is widely used in face recognition studies (Wei et al., 2020; Morozov et al., 2021). For each species, we generated an aligned and an unaligned dataset. Adapting this approach to our

case study, we used a custom-made Python script to compute the angle of rotation between the coordinates of the two inner eye parts (RightEye_Inner-LeftEye_Inner) and to rotate the landmarks, ensuring that the line connecting these key points was horizontal (180°).

We first calculated the Euclidean distance between each pair of the 13 points, resulting in a 13x13 distance matrix for each frame. To account for the variability due to the animal's distance from the recording camera, we normalised all the matrices using the distance between the RightEye_Inner and LeftEye_Inner points, which remains fixed at the individual level and constant regardless of facial gestures (following Zhang et al., 2016). This helped to mitigate the discrepancies derived from the subject distance during recordings. After normalising the matrices, we imported them into the R software (R Core Team 2020, version 2023.12.1+402). We tabulated them to construct a data frame consisting of 78 variables corresponding to the number of non-redundant or constant (i.e., 0 and 1) distances.

Figure 1 summarises the composition of the key points set, model development, and data processing/extraction.

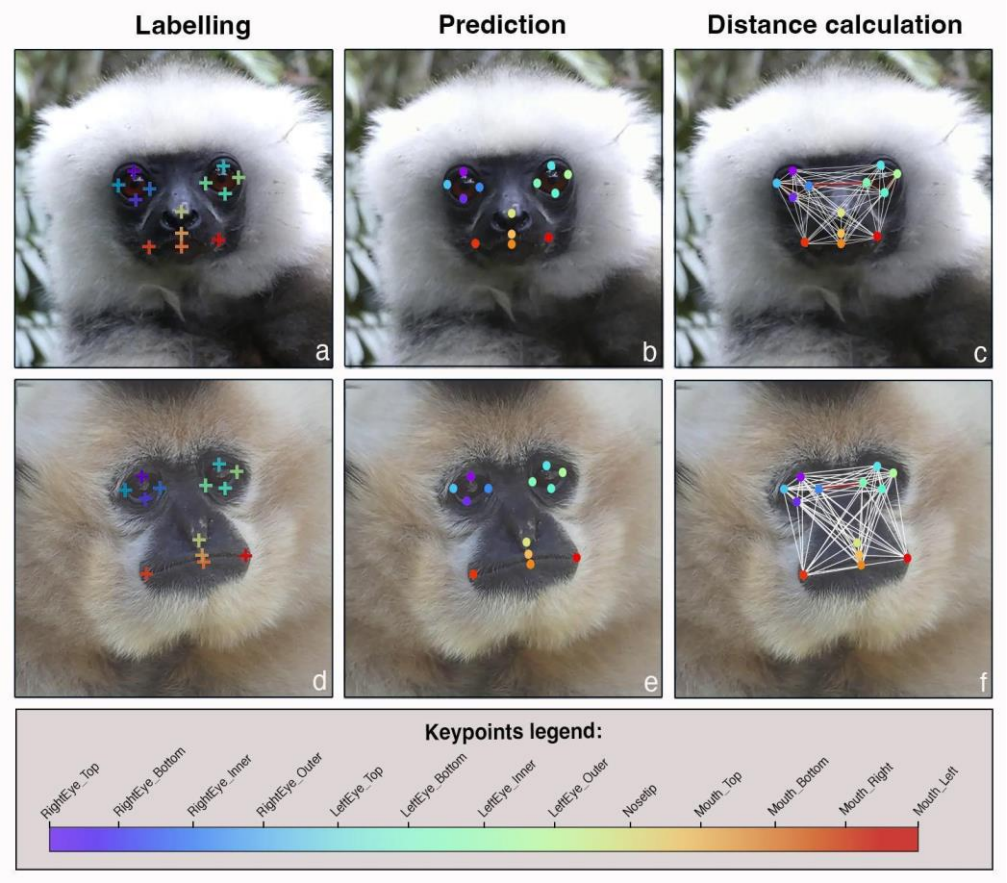


Figure 1. The set of key points used for training the model on DLC: labelled by the human operator (a-d), predicted by the DLC model (b-e), converted in distance matrix, and normalised (the red line shows the distance selected for normalisation) (c-f).

Since our resulting set presents numerous highly correlated variables, we used two approaches to prepare the data for subsequent automatic classifications. In the first case, we performed a correlation analysis (R Package stats, version 4.1.2) and removed all variables more correlated than 0.75 from the datasets. The variables selected for each dataset are listed in the Supplementary Materials (S1_List_variables_for_each_ML.docx). A second approach was to apply a principal component analysis (PCA) (R package FactoMineR; Husson et al., 2016) for each target species

to reduce data to new uncorrelated variables (PCs). Then, we selected only the components with eigenvalue >1 (ten for *I. indri*, seven for *P. diadema* and nine for *N. gabriellae*). Results of Principal Component Analysis on aligned and unaligned data are reported in the supplementary materials (S2_PCA_supplementary_results.doc).

We then approached the classification process for each species using four datasets with different preprocessing treatments: i) PCA on unaligned data, ii) uncorrelated variables from the unaligned data, iii) PCA on aligned data, iv) uncorrelated variables from the aligned data. This multiple datasets approach allowed us to understand how the individual treatments affected the algorithms' classification capabilities.

2.4 Classification algorithms

To assess the ability to distinguish between voiced and unvoiced facial configurations, we employed three machine learning algorithms: i) a multi-layer perceptron (referred to as MLP), ii) a support vector machine (SVM), and iii) a random forest classifier (RFC). Since the frames in which the animals were vocalising accounted for only about 10% of those they were not, we selected an equal number (N = 300) of instances for each class in every run-through subsampling. We executed each algorithm 100 times and ran the classification process for each pre-processed dataset.

For the Multilayer Perceptron (MLP), we used the *mlp* function from the RSNNS package with learnFuncParams set to 0.1 and maxit set to 100 (Bergmeir & Benitez-Sánchez, 2012). For the Support Vector Machine (SVM), we used the *SVM* function from the e1071 package. We tested

gamma values of 0.005, 0.010, 0.015, 0.020, 0.025, 0.030, 0.035, 0.040, 0.045, and 0.050 for tuning, as well as cost values of 10^-8, 10^-4, 10^-2, and 10^0. We also tuned coef0 with values of 0.1, 1, and 10. We utilised C-classification and a polynomial kernel with a degree of 2 (Dimitriadou et al., 2006). We selected the best gamma, cost, and coef0 values from tuning to achieve the highest classification rates. For the Random Forest Classifier (RFC), we used the *randomForest* function from the randomForest package in R with N trees set to 500 and N variables at each split set to 3. We then trained each classifier using 70% of each subsample and tested it on the remaining 30%. We calculated the average correct classification rates and their standard deviation for MLP, SVM, and RFC. After checking the distribution of the correct classification rates using the Shapiro-Wilk test (Shapiro & Wilk, 1965), we tested for significant differences in the correct classification rates using the Paired t-test (De Winter, 2019). In cases of deviation from normal distribution, we used the Wilcoxon Paired Test from the coin package (Hothorn et al., 2008). To evaluate the effect of the different preprocessing steps on the classification of the best-performing algorithm, we compared the correct classification rates using Student t-tests (Student, 1908) and Mann-Whitney U tests (McKnight and Najab, 2010).

Shapley coefficients efficiently explain what happens to a model when we change the value of features and provide consistent insight into which features have the greatest influence in a machine-learning process (Strumbelj and Kononenko, 2014). For the sake of clarity, we only report SHAP analysis results for Random Forest classifications.

3. RESULTS

3.1 DLC models performance

We developed three DLC models capable of efficiently identifying the positions of our landmark set, demonstrating low root mean square error (RMSE) across all examined species: 3.72 px (with p-cutoff: 2.78 px) for *I. indri*, 3.29 px (with p-cutoff: 3.12 px) for *P. diadema* and 4.96 px (with p-cutoff: 4.12 px) for *N. gabriellae*. Figure S3 shows root mean square error variation with the number of iterations in training and testing, for the three species. We provided readers with examples of labelled videos using the developed models in the supplementary material (Videos S4–S6).

The analysis of novel videos indicated that the developed DLC models could generalise to clips not used for sampling images in the training and testing sets. The Mean Absolute Euclidean Distance (MEAD) between manually labelled and predicted key points was 5.68 ± 8.06 px for *I. indri*, 6.61 ± 13.70 px for *P. diadema* and 6.17 ± 1.20 px for *N. gabriellae.* As summarised in Table 2, the MEAD values revealed consistent differences among each key point, with the lowest performance observed in the peripheral parts of the mouth (Mouth_Right and Mouth_Left), and this trend is shared across all the target species. Examples of labelled novel videos are reported in the supplementary material (Videos S7–S9).

Key-points	*I.indri*		*P. diadema*		*N. gabriellae*	
	MEAD (SD)	N detections	MEAD (SD)	N detections	MEAD (SD)	N detections
RightEye_top	5.37 (19.49)	107	2.76 (1.81)	143	4.68 (1.20)	106
RightEye_Bottom	4.30 (2.87)	107	4.52 (2.81)	145	5.25 (3.21)	111
RightEye_Inner	3.96 (2.93)	93	5.65 (12.34)	136	6.31 (2.25)	96
RightEye_Outer	6.68 (25.03)	106	8.85 (46.55)	141	5.84 (3.01)	104
LeftEye_top	4.86 (4.12)	123	5.19 (20.58)	132	6.77 (1.69)	108
LeftEye_Bottom	5.00 (3.70)	111	4.21 (2.54)	132	7.48 (3.58)	110
LeftEye_Inner	5.17 (3.46)	92	5.11 (3.76)	108	6.65 (1.00)	90
LeftEye_Outer	4.22 (2.72)	120	9.92 (36.6)	127	5.70 (2.71)	105
Nosetip	6.53 (21.08)	158	5.75 (4.43)	149	4.17 (3.30)	125
Mouth_Top	5.31 (3.86)	86	7.64 (12.76)	127	6.69 (2.37)	114
Mouth_Bottom	5.13 (3.13)	78	6.44 (7.91)	111	5.31 (2.16)	93
Mouth_Right	7.97 (4.53)	29	10.68 (11.99)	53	7.29 (2.07)	81
Mouth_Left	9.30 (7.90)	40	10.56 (12.76)	36	8.03 (6.59)	82

Table 2. Model performance across facial key points within novel videos. The number of detections that overcome the p-cutoff is reported for each landmark. MEAD = mean absolute Euclidean distance.

3.2 Classification results

All the algorithms showed a correct classification rate of vocalising and non-vocalising facial gestures, which was higher than the chance for all species, and all the datasets were pre-processed differently (Figure 2). Overall, the algorithm that performed the best classification was RFC, with the highest results when applied to uncorrelated distances calculated from either aligned or not coordinates. However, we detected slight differences in the algorithm performances between species. The mean correct classification rates (and standard deviation) for all ML algorithms and the mean and standard deviation of the other metrics for the best performing dataset (correct classification rate, precision, recall, F1, and AUC score) are reported in the supplementary materials (Tables S10).

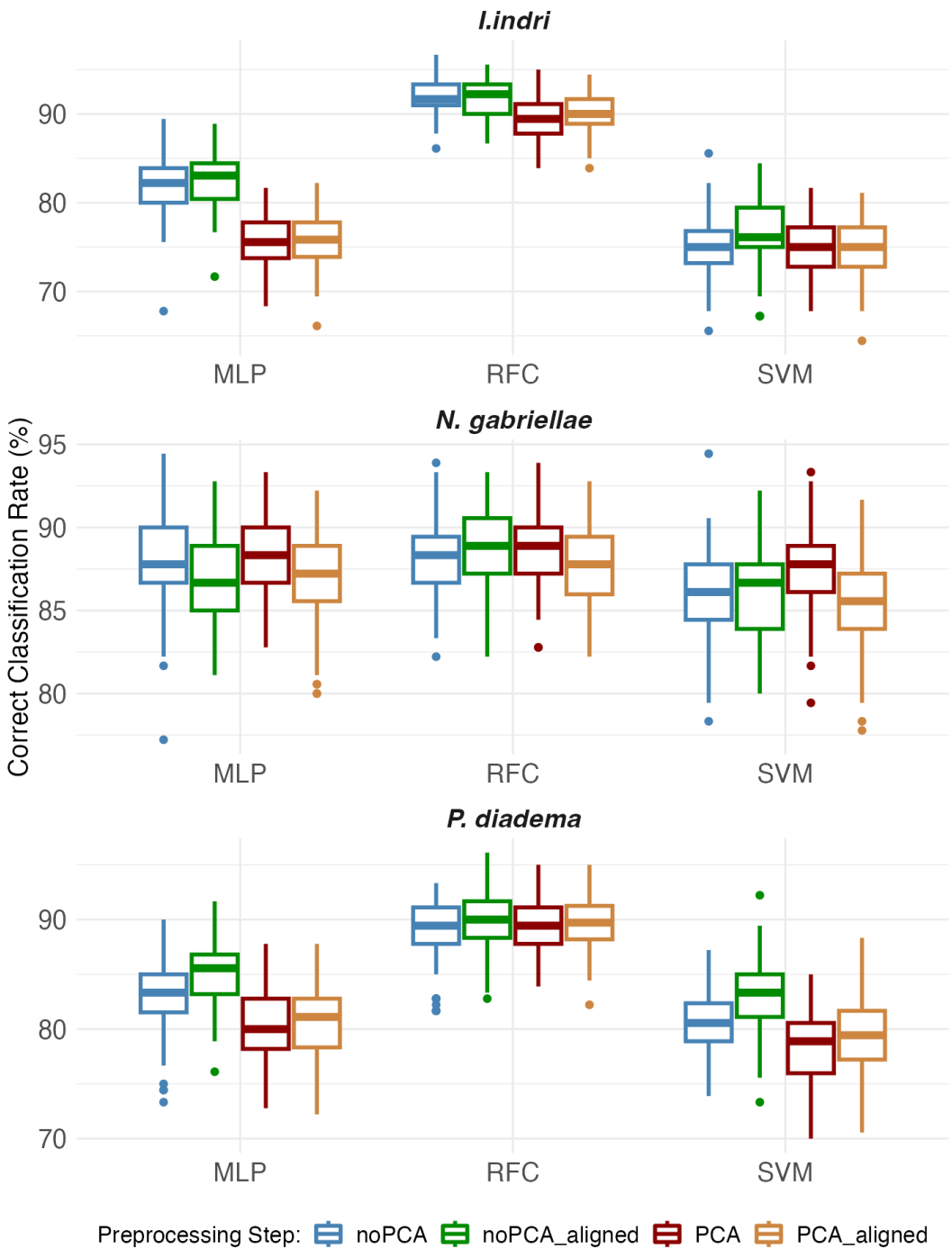


Figure 2. Boxplots show the correct classification rates for each machine learning algorithm: multilayer perceptron (MLP), random forest classifier (RFC), and support vector machine (SVM). Each subfigure refers to a different

primate species: *Indri indri* (A), *Nomascus gabriellae* (B), and *Propithecus diadema* (C). Colours denote different processing and preprocessing steps.

Starting from *I. indri*, all algorithms showed high values of correct classification rates (Figure 3). However, RFC outperformed other machine-learning techniques across all pre-processed datasets (Figure 2). RFC achieved the highest performance on the dataset derived from unaligned coordinates (92.01 $\pm$ 1.99 %; Table 3). Wilcoxon tests revealed significant differences among the techniques applied (RFC-SVM: $V = 0$, $p < 0.01$; MLP-RFC: $V = 0$, $p < 0.01$). Comparison of the RFC correct classification rates across different datasets revealed significant differences between those treated with principal component analysis (Mann-Whitney test; noPCA vs PCA: $W = 7943$, $p < 0.01$; noPCA_aligned vs PCA_aligned: $W = 8191$, $p < 0.01$), while no significant differences emerged when comparing aligned and non-aligned data without PCA (Mann-Whitney test; noPCA vs noPCA_aligned: $W = 4252$, $p = 0.39$).

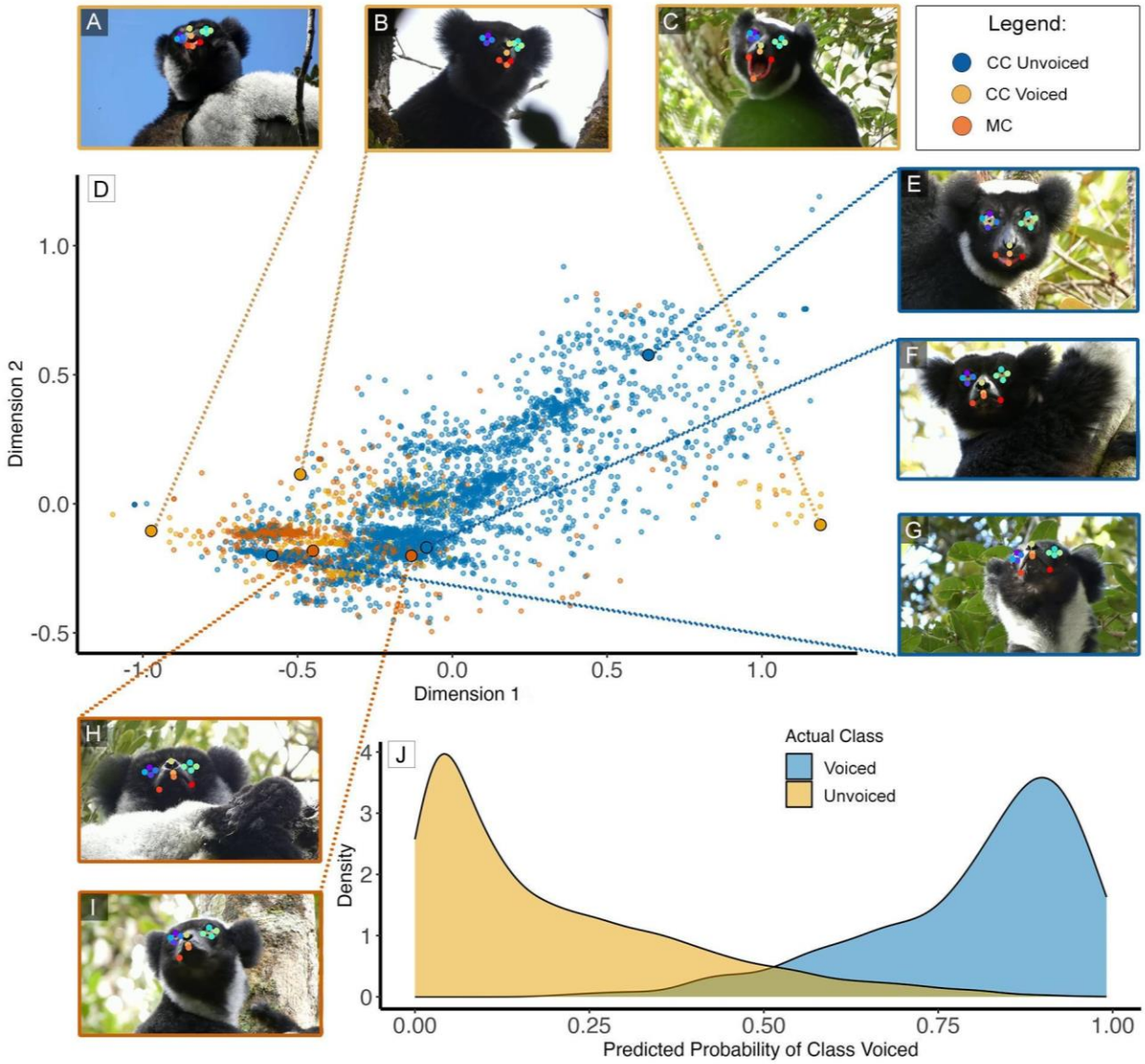


Figure 3. Random Forest classification results (D, J) on *I.indri* key point distances derived from unaligned coordinates (CC = correct classification, MC = misclassification). We used multidimensional scaling (MDS) to explore the classification results in a bi-dimensional space identifying a frame close to the 1st quartile for Dimension 1 for voiced gestures (C) and unvoiced gestures (E), the 2nd quartile (B and F respectively), and the 3rd quartile (A and G). Figures H and I display two misclassified gestures. On the lower right corner, a density plot (J) showing the predicted probability to the class "Voiced" (i.e. voiced gestures)

We obtained similar results for *P. diadema* (Figure 4): all the algorithms showed good classification performance, with RFC as best when applied on distances computed from aligned coordinates (89.85 ± 2.81 %; Table 3). As in the previous case, the paired t-tests revealed significant differences among the methodologies (RFC-SVM: $T = -23.93$, $p < 0.01$; MLP-RFC: $T = 10.90$, $p < 0.01$). Unlike *I. indri*, no significant differences emerged from the comparison with dimensionally reduced data (Mann-Whitney test; noPCA_aligned vs PCA: $W = 5508$, $p = 0.21$; noPCA_aligned vs PCA_aligned: $W = 5281$, $p = 0.49$), while not aligned data showed significantly

lower performances (Mann-Whitney test; noPCA vs noPCA_aligned: W = 5955, p < 0.05).

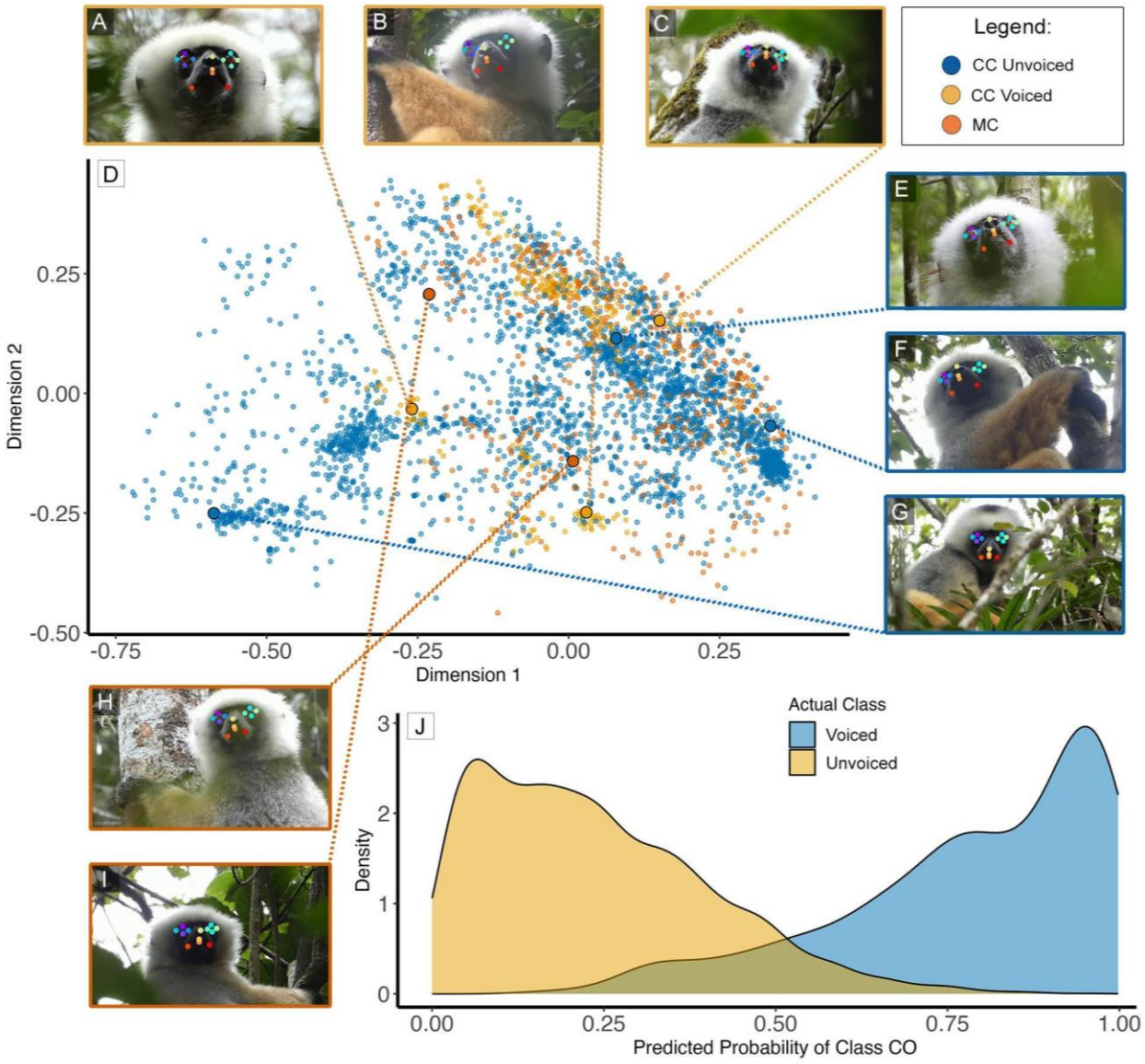


Figure 4. Random Forest classification results (D, J) on *P. diadema* key point distances derived from aligned coordinates (CC = correct classification, MC = misclassification). We used multidimensional scaling (MDS) to explore the classification results in a bi-dimensional space identifying a frame close to the 1st quartile for Dimension 1 for voiced gestures (C) and unvoiced gestures (F), the 2nd quartile (B and E respectively), and the 3rd quartile (A and G). Figures H and I display two misclassified gestures. On the lower right corner, a density plot (J) showing the predicted probability to the class "Voiced" (i.e. voiced gestures)

Concerning *N. gabriellae*, the analysis showed very few differences among the classification rates of the three techniques. However, the best performance is again represented by RFC (Figure 5) applied on aligned data (88.94 ± 2.18; Table 3), and the paired t-test revealed significant differences between RFC and the other techniques (MLP-RFC: $T = -6.11$ $p < 0.01$; SVM-RFC: $T = -14.40$, $p < 0.01$). Performance of aligned data was significantly higher than not aligned data (t-test; noPCA_aligned vs noPCA: $T = 2.17$, $p < 0.05$) and than dimensionally reduced aligned data (t-test; noPCA_aligned vs PCA_aligned: $T = 3.81$, $p < 0.01$), while we found no differences from the comparison with not aligned principal components (t-test; noPCA_aligned vs PCA: $T = 0.89$, $p = 0.37$).

The SHAP analysis indicated how strongly the individual variables contributed to model prediction. We understand from the average SHAP values that Mouth_Left-Mouth_Top, LeftEye_Outer-LeftEye_Bottom, Nosetip-RightEye_Bottom, were the most critical distances for the indris in determining voiced facial gestures. LeftEye_Outer-LeftEye_Bottom and Mouth_Right-Mouth_Bottom, and LeftEye_Inner-RightEye_Top and Mouth_Left-Mouth_Right were the most important in models of *Nomascus gabriellae* and *Propithecus diadema*, respectively (Figure S11).

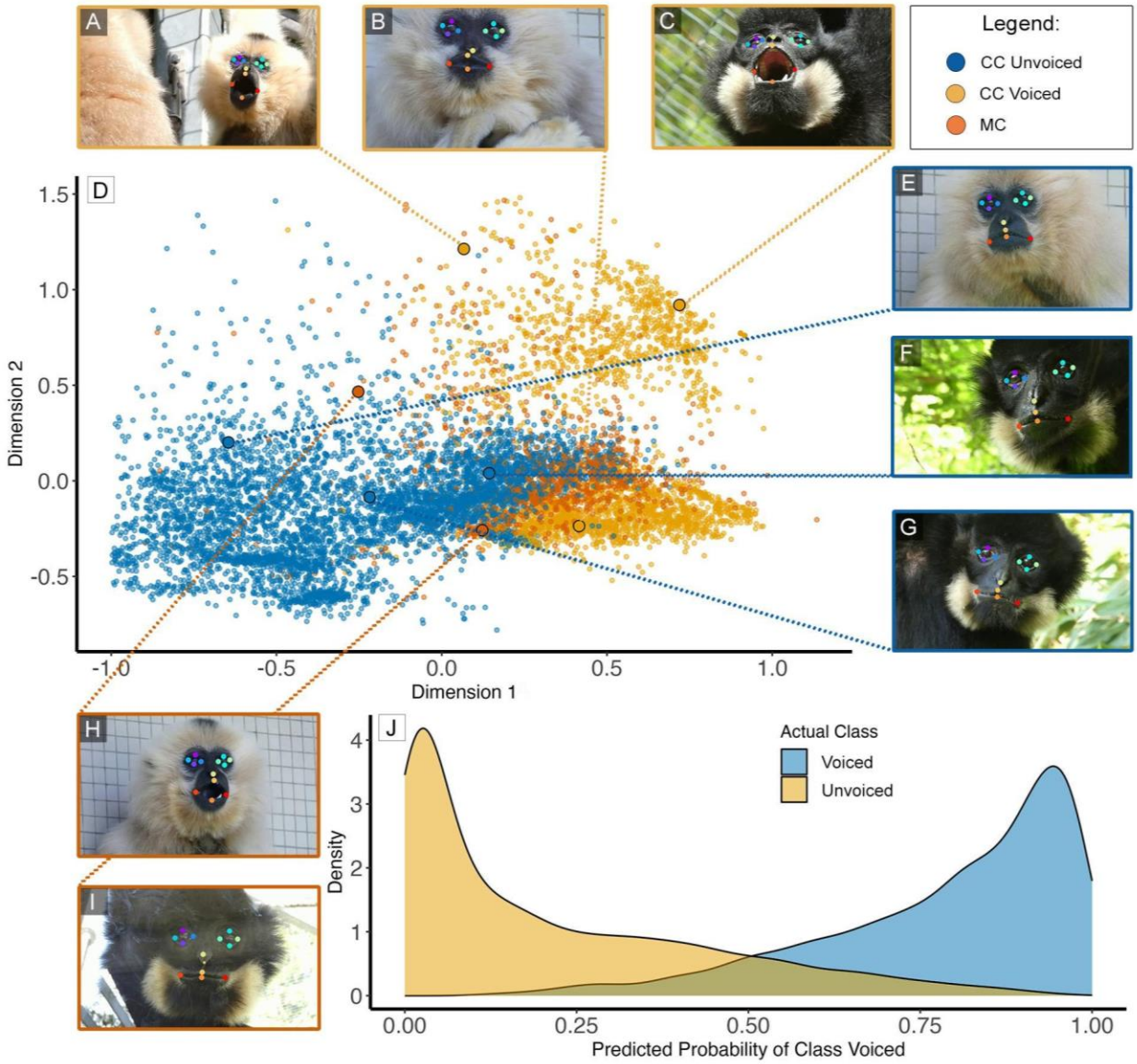


Figure 5. Random Forest classification results (D, J) on *N. gabriellae.* Key-point distances derived from aligned coordinates (CC = correct classification, MC = misclassification). We used multidimensional scaling (MDS) to explore the classification results in a bi-dimensional space identifying a frame close to the 1st quartile for Dimension 1 for voiced gestures (C) and unvoiced gestures (F), the 2nd quartile (B and G respectively), and the 3rd quartile (A and E). Figures H and I display two misclassified gestures. On the lower right corner, a density plot (J) showing the predicted probability to the class "Voiced" (i.e. voiced gestures)

4. DISCUSSION

We used advanced deep-learning algorithms to analyse the facial expressions of primate species. Our research demonstrates that facial gestures associated or not with vocalisations can be distinguished using distances between 13 key points. High correct classification rates are held across algorithms and in both captive and wild conditions.

1. Deep learning application for studying facial gestures

Our findings extend previous applications of DeepLabCut for studying animal behaviour by predicting the position of facial landmarks in different primate species. Our DLC models showed RSME for facial landmarks lower than 5 pixels, which is in line with the existing models considering animal-specific body parts (Sato et al., 2022; Wrench and Balch-Tomes, 2022). Results from our study align with previous DLC models developed for primate pose estimation, such as MacaquePose, based on captive macaque recordings (Labuguen et al., 2019; Labuguen et al., 2021), and DeepWild, built from wild chimpanzees and bonobos videos (Wiltshire et al., 2023), confirming that key points positioned on the subject face are detected efficiently. This result corroborates the suitability of DLC application in studying primate facial configurations. Despite multi-animal design and a larger key-point set to track the entire body, facial landmarks used in those applications showed lower RMSE and better performance than the other body points with error rates similar to our study (Labuguen et al., 2019; Labuguen et al., 2021; Wiltshire et al., 2023).

We observed differences in the performances among the three species. Lemurs' models performed better (i.e. lower RMSE) than the crested gibbon model. This result is surprising given that the gibbon model was trained on more images, taken in captivity, and included fewer individuals. These factors would presumably be associated with lower variability from the recording environment. A possible explanation for these differences in performance can reside in the species-specific morphological characteristics. Unlike indris and sifakas, yellow-cheeked crested gibbons have highly dimorphic and dichromatic face and head fur patterns, showing a higher degree of inter-individual variation (Bolechovà et al., 2016; Mootnick and Fan, 2011).

Our novel video analysis tested the ability to generalise to new footage of the DLC models, including different recording conditions and previously unseen individuals. The results showed performance errors (i.e., MEAD values) higher than each model's test errors but remarkably lower than those reported by Wiltshire and colleagues (2023), supporting the robustness of our models in identifying our set of facial landmarks and their suitability for the analysis of new recordings. We also found differences in the MEAD across each landmark, with higher errors and lower detection rates in distal points such as the mouth's (Mouth_right and Mouth_left) and eyes' sides (Right_Eye_Outer and Left_Eye_Outer).

Considering the facial morphology of the study species (e.g., lemurs' facial protrusion), we can understand how the orientation of the face with respect to the camera is an element that can easily make some key points less visible or more challenging to predict. This finding aligns with differences among landmarks observed in the DeepWild novel video analysis (Wiltshire et al.,

2023), supporting that each key point's position and intrinsic features can influence its detectability (e.g. landmarks in highly contrasted areas are more easily identifiable than those close to fur-covered regions).

2. Pre-processing steps and classification results

We used different processing approaches to understand those performing best after calculating distances between key points and normalising them to the interocular distance (Zhang et al., 2016). We tested facial alignment and PCA's influence on the further classification results. The comparison revealed that facial alignment is important in improving the classification accuracy of all machine-learning techniques (Zhang et al., 2016). On the contrary, principal component analysis, applied to aligned or unaligned data, decreased classification accuracy. Applying PCA before discrimination is often debated when there is a relatively large number of variables and just a few levels of the grouping factor. In our case, variable reduction (i.e. removing highly correlated variables) proved more efficient than PCA, leading to higher rates across all the machine-learning approaches. Our findings align with previous applications of machine learning in ecology, where classification accuracy improved when using a small set of selected features (Tirelli & Pessani, 2011; Tirelli et al., 2011) but contrast results of studies in which ML classification improved following the use of data reduction by PCA (e.g., Awan et al., 2019). Evidence shows that using principal components can negatively impact decision tree-based techniques, including random forests, which performed exceptionally well in our study (Howley et al., 2006). While, on the one hand, our results corroborate the fact that PCA is not to be used when dealing with Decision Trees or Random Forests, on the other hand, we have to record how some studies have screened out the

effect of adding principal components to the original parameters during a classification process. We see this as a prospect that can be evaluated in later studies (Popelinsky & Brazdil, 2000).

An important dogma in ML is that one single algorithm might not necessarily be the best across all possible classification problems (Boateng et al. 2020) but, rather, the efficacy depends on type and dimensionality of datasets, as well as according to the measure employed for comparing classifiers (Gupta et al. 2022). Nonetheless, it seems that Random Forest outperforms a wide range of other methods and is one of the most popular algorithms for various prediction and classification tasks (Sheykhmousa et al. 2020). In fact, out of 68 studies published between 2000 and 2017, most recommend SVM and RFC because of higher accuracy and easier implementation (Boateng et al. 2020). RFC has been demonstrated to be the best classifier in both training and testing phase of financial risk evaluation (Dong et al. 2024), in disease prediction (reaching the highest accuracy in 9 studies out of 17 where it was employed: Uddin et al. 2019), in land use classification (Ramachandra et al. 2023), and across 121 datasets belonging to the UCI ML repository (achieving the best performance in more than 90% of cases: Fernández-Delgado et al. 2014). In line with these studies, we found RFC to outperform both SVM and MLP in all classification tasks. Furthermore, unlike Gupta and colleagues (2022), but in line with Chowdhury 2024 (employing ML algorithms to demarcate built up and bare land in urban settings), we found that RF was the best classifier regardless of the measure employed for evaluation (see table S8). Again in line with Chowdhury 2024, we also found that MLP performed slightly better than SVM. Hence, our results represent a further shred of evidence sustaining RFC to be one of the easier, faster, and more accurate algorithms available. Provided that the reasons for performance lie in the data rather than in the algorithms themselves (Gupta et al. 2022), RF seems to be more efficient in treating large

input datasets and to perform better in mixed classes classification than SVM (Adugna et al. 2022). Our configurations, being the result of subtle facial movements, might be intrinsically mixed, and might have caused the SVM to be more prone to confusion among classes. Moreover, Neural Networks are known to need more data than RF to achieve a comparable accuracy (Roßbach 2018). See Boateng et al. 2020 and Roßbach 2018 for thorough comparisons highlighting weaknesses and strengths of various ML algorithms.

3. Voiced vs Unvoiced Classification results

Supervised machine-learning techniques showed high correct classification rates for efficient discrimination between vocalised and non-vocalised facial gestures. Despite the fewer landmarks compared to the model developed by Witham (2018), our custom set of key points can efficiently summarise face configurations, highlighting the differences between faces during the emission of vocalisation or in unvoiced gestures.

Our results indicate that our approach can successfully screen a large set of videos and pictures to target voiced gestures. These gestures are indeed of interest in a wide range of comparative studies, from the multimodal nature of primate communication to the evolution of vocal communication in the animal kingdom (Ghazanfar & Takahashi, 2014). Our methodology could reduce the mismatch between the high volume of raw data often collected in ecological studies and the ability to extract meaningful information with little human labour (Tuia et al., 2022).

Our findings agree with studies that have shown that vocal emission modifies facial appearance in humans (Lyons et al., 1998; Hontanilla & Aubá, 2008; Dagnes et al., 2019; Yehia et al., 1998; Yehia et al., 2002) and non-human primates (rhesus macaques - Hauser et al., 1993; Hauser & Ybarra, 1994; Ghazanfar, 2013). Oral tract movements during vocalisation emission can determine

remarkable modifications in face configuration. For instance, rhesus macaques can change their lip configuration according to the emitted vocal types, protruding their lips while emitting a coo call or retracting them during a scream (Hauser et al., 1993; Hauser & Ybarra, 1994). The co-occurrence of vocalisations and face changes have also been described in lemurs, as shown in black-and-white ruffed lemurs roar-shriek chorus (Gamba & Giacoma, 2006) or indri songs (*Indri indri*), which modulate the mouth configuration while singing (Favaro et al., 2008; Gamba et al., 2011). However, particular call types can be emitted with the mouth closed (e.g. *hum* and *mmm* in *Propithecus diadema* - see Valente et al. 2022) or barely open (e.g. *grunt* in *Indri indri* - see Maretti et al. 2010). Therefore, future research could investigate whether nasal resonating calls correspond to subtle variations in the facial configuration not captured by the current key points, potentially resulting in misclassification.

The visual inspection of misclassified frames provided insights into the potential causes behind the incorrect classifications, revealing various scenarios. Most misclassified cases were "unvoiced" gestures predicted as "voiced". In several instances, these frames depicted subjects with open mouths, such as during chewing, supporting the idea that mouth opening plays a critical role in vocal emissions (Fitch et al., 2016). Another factor likely influencing misclassification was the camera angle. Many misclassified unvoiced configurations, particularly in the indri model, showed subjects with their heads rotated upward. Since head rotation characterises the emission of indris' song (Gamba et al., 2011), the most represented vocal type within the present study, the camera angle may have contributed to the misclassification of images framing subjects with heads rotated upward. Fewer instances involved voiced gestures incorrectly classified, reinforcing that facial configuration is highly distinctive during vocal emission in the three study species. Another potential source of inconsistency in classification is that, since we were working at a frame-level

analysis, the clips, including vocal emissions, could include gestures immediately anticipating or following vocalisation. Those gestures could be incorrectly labelled as "voiced" and then misclassified. Thus, if labelling frames begins at the video screening stage, the precision with which the videos are cut from the footage is important for generating frames correctly labelled a priori.

Interestingly, SHAP analyses showed that variables related to oral movements in all species were highly important for classification of voiced facial configurations. This results supports the idea that jaws movements are critical in determining distinctive facial gestures in nonhuman primates (Ghazanfar, 2013).

CONCLUSIONS

In the future, we could use deep learning to decode the facial movements of other animal species. Our study has shown that this potential exists, and with increasingly powerful computers, this process could become even more accessible. However, this does not mean that human screening of images is unnecessary, as it currently allows the detection of differences that can be encoded in messages exchanged through facial gestures. Further studies could investigate the ability of DL algorithms to identify and classify different facial gestures. This approach could help integrate with existing techniques like AnimalFACS (Ekman & Frieser, 1978; Waller et al., 2020) to contribute to developing comparative studies concerning communicative multimodality.

REFERENCES

Adugna, T., Xu, W., Fan, J. (2022). Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images. Remote Sensing, 14(3), 574. https://doi.org/10.3390/rs14030574

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., Chow, B. J., Dwivedi, G. (2019). Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. PloS one, 14(6), e0218760. https://doi.org/10.1371/journal.pone.0218760

Bala, P. C., Eisenreich, B. R., Yoo, S. B. M., Hayden, B. Y., Park, H. S., Zimmermann, J. (2020). Automated Markerless Pose Estimation in Freely Moving Macaques with OpenMonkeyStudio. Nature Communication, 11(1):4560. https://doi.org/10.1038/s41467-020-18441-5;

Bergmeir, C. N., Benitez-Sánchez, J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software, 46(7), ISSN: 1548-7660. https://doi.org/10.18637/jss.v046.i07;

Boateng, E., Otoo, J. Abaye, D. (2020) Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. Journal of Data Analysis and Information Processing, 8, 341-357. doi: 10.4236/jdaip.2020.84020.

Bolechová, P., Ječmínková, K., Hradec, M., Kott, T., & Doležalová, J. (2017). Sex determination in gibbons of genus Nomascus using non-invasive method. Acta Veterinaria Brno, 85(4), 363-366. https://doi.org/10.2754/avb201685040363

Chowdhury, M. S. (2024). Comparison of accuracy and reliability of random forest, support vector machine, artificial neural network and maximum likelihood method in land use/cover classification of urban setting. Environmental Challenges, 14, 100800. https://doi.org/10.1016/j.envc.2023.100800

Cohn, J. F., Zlochower, A. J., Lien, J., Kanade, T. (1999). Automated Face Analysis by Feature Point Tracking has High Concurrent Validity with Manual FACS Coding. Psychophysiology, 36(1), 35-43. https://doi.org/10.1017/S0048577299971184;

Crouse, D., Jacobs, R. L., Richardson, Z., Klum, S., Jain, A., Baden, A. L., Tecot, S. R. (2017). LemurFaceID: A Face Recognition System to Facilitate Individual Identification of Lemurs. Bmc Zoology, 2(1):2. https://doi.org/10.1186/s40850-016-0011-9;

Dagnes, N., Marcolin, F., Vezzetti, E., Sarhan, F. R., Dakpé, S., Marin, F., Nonis, F., Mansour, K. B. (2019). Optimal Marker Set Assessment for Motion Capture of 3D Mimic Facial Movements. Journal of Biomechanics, 93, 86-93. https://doi.org/10.1016/j.jbiomech.2019.06.012;

De Winter, J. C. (2019). Using the Student's T-test with Extremely Small Sample Sizes. Practical Assessment, Research, and Evaluation, 18(10). https://doi.org/10.7275/e4r6-dj05;

Deb, D., Wiper, S., Gong, S., Shi, Y., Tymoszek, C., Fletcher, A., Jain, A. K. (2018). Face recognition: Primates in the Wild. In: IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), 1-10. https://doi.org/10.1109/BTAS.2018.8698538;

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., Leisch, M. F. (2006). The e1071 Package. Misc Functions of Department of Statistics (e1071). TU Wien, 297-304. Google Scholar;

Dong, H., Liu, R., Tham, A. W. (2024). Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation. Journal of Risk and Financial Management, 17(2), 50. https://doi.org/10.3390/jrfm17020050

Dufourq, E., Durbach, I., Hansford, J. P., Hoepfner, A., Ma, H., Bryant, J. V., Stender, C. S., Li, W., Liu, Z., Chen, Q., Zhou, Z., Turvey, S. T. (2021). Automated Detection of Hainan Gibbon Calls for Passive Acoustic Monitoring. Remote Sensing in Ecology and Conservation, 7(3), 475-487. https://doi.org/10.1002/rse2.201;

Ekman, P., Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychology Journal, Palo Alto. Google Scholar;

Favaro, L., Gamba, M., Sorrentino, V., Torti, V., Giacoma, C. (2008). Singers in the Forest: Acoustic Structure of Indri's Loud Calls and Vocal Tract Tuning in a Prosimian Primate. Rivista Italiana Acustica, 32, 35. Google Scholar;

Feilghelstein, M., Shimshoni, I., Finka, L. R., Luna, S. P., Mills, D. S., Zamansky, A. (2022). Automated recognition of pain in cats. Scientific Reports, 12(1), 9575. https://doi.org/10.1038/s41598-022-13348-1

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The journal of machine learning research, 15(1), 3133-3181.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. The Journal of the Acoustical Society of America, 102(2), 1213-1222. https://doi.org/10.1121/1.421048

Fitch, W. T., De Boer, B., Mathur, N., & Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. Science advances, 2(12), e1600723. https://doi.org/10.1126/sciadv.1600723

Friard, O., Gamba, M., (2016). BORIS: a Free, Versatile Open-Source Event-Logging Software for Video/Audio Coding and Live Observations. Methods in Ecology and Evolution, 7(11), 1325-1330. 10.1111/2041-210X.12584;

Fuchs, M., Genty, E., Zuberbuhler, K., Cotofrei, P. (2023). ASBAR: an Animal Skeleton-Based Action Recognition framework. Recognizing great ape behavior in the wild using pose estimation with domain adaptation. BioRxiv (Bio-Archive), The preprint server for biology. https://doi.org/10.1101/2023.09.24.559236;

Gamba, M., Favaro, L., Torti, V., Sorrentino, V., Giacoma, C. (2011). Vocal Tract Flexibility and Variation in the Vocal Output in Wild Indris. International Journal Animal Sound and its Recording, 20(3), 251-265. https://doi.org/10.1080/09524622.2011.9753649;

Gamba, M., Friard, O., Riondato, I., Righini, R., Colombo, C., Miaretsoa, L., Torti, V., Nadhurou, B., Giacoma, C. (2015). Comparative Analysis of the Vocal Repertoire of Eulemur: A Dynamic Time Warping Approach. International Journal of Primatology, 36(5), 894-910. https://doi.org/10.1007/s10764-015-9861-1;

Gamba, M., Giacoma, C. (2006). Vocal Tract Modeling in a Prosimian Primate: The Black and White Ruffed Lemur. Acta Acustica United with Acustica, 92(5), 749-755, Acta Acustica;

Ghazanfar, A. A. (2013). Multisensory Vocal Communication in Primates and the Evolution of Rhythmic Speech. Behavioral Ecology and Sociobiology, 67, 1441-1448. https://doi.org/10.1007/s00265-013-1491-z;

Ghazanfar, A. A., Takahashi, D. Y. (2014). Facial expressions and the evolution of the speech rhythm. Journal of cognitive neuroscience, 26(6), 1196-1207. https://doi.org/10.1162/jocn_a_00575

Guo, S., Xu, P., Miao, Q., Shao, G., Chapman, C. A., Chen, X., He, G., Fang, D., Zhang, H., Sun, Y., Shi, Z., Li, B. (2020). Automatic Identification of Individual Primates with Deep Learning Techniques. iScience, 23(8):101412. https://doi.org/10.1016/j.isci.2020.101412;

Gupta, S., Saluja, K., Goyal, A., Vajpayee, A., Tiwari, V. (2022). Comparing the performance of machine learning algorithms using estimated accuracy. Measurement: Sensors, 24, 100432. https://doi.org/10.1016/j.measen.2022.100432

Hamm, J., Kohler, C. G., Gur, R. C., Verma, R. (2011). Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. Journal of neuroscience methods, 200(2), 237-256. https://doi.org/10.1016/j.jneumeth.2011.06.023

Hardin, A., Schlupp, I. (2022). Using Machine Learning and DeepLabCut in Animal Behavior. Acta Ethologica, 25, 125-133. https://doi.org/10.1007/s10211-022-00397-y;

Hauser, M. D., Evans, C. S., Marler, P. (1993). The Role of Articulation in the Production of Rhesus Monkey, *Macaca mulatta*, vocalizations. Animal Behaviour, 45(3), 423-433. https://doi.org/10.1006/anbe.1993.1054;

Hauser, M. D., Ybarra, M. S. (1994). The Role of Lip Configuration in Monkey Vocalizations: Experiments Using Xylocaine as a Nerve Block. Brain and Language, 46(2), 232-244. https://doi.org/10.1006/brln.1994.1014;

Hayakawa, S., Kataoka, K., Yamamoto M., Asahi T., Suzuki, T. (2024). DeepLabCut - based daily behavioural and posture analysis in a cricket. Biology Open, 13(4). https://doi.org/10.1242/bio.060237;

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the Institute of Electrical and Electronics Engineers (IEEE), 770-778. https://doi.org/10.48550/arXiv.1512.03385;

Hontanilla, B., Aubá, C. (2008). Automatic Three-Dimensional Quantitative Analysis for Evaluation of Facial Movement. Journal of Plastic, Reconstructive & Aesthetic Surgery, 61(1), 18-30. https://doi.org/10.1016/j.bjps.2007.03.037;

Hothorn, T., Hornik, K., Van De Wiel, M. A., Zeileis, A. (2008). Implementing a Class of Permutation Tests: the Coin Package. Journal of Statistical Software, 28(8), 1-23. https://doi.org/10.18637/jss.v028.i08;

Howley, T., Madden, M. G., O'Connell, M. L., Ryder, A. G. (2005). The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence. London: Springer London, pp. 209-222.

Husson, F., Josse, J., Le, S., Mazet, J., & Husson, M. F. (2016). Package 'factominer'. An R package, 96(96), 698.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B. (2016). DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In B. Leibe, J. Matas, N. Sebe,

M. Welling, European Conference on Computer Vision ECCV, 9910, 35-40, Springer. https://doi.org/10.1007/978-3-319-46466-4_3;

Janisch, J., Mitoyen, C., Perinot, E., Spezie, G., Fusani, L., Quigley, C. (2021). Video Recording and Analysis of Avian Movements and Behavior: Insights from Courtship Case Studies. Integrative and Comparative Biology, 61(4), 1378-1393. https://doi.org/10.1093/icb/icab095;

Kaminski, J., Waller, B. M., Diogo, R., Hartstone-Rose, A., Burrows, A. M. (2019). Evolution of Facial Muscle Anatomy in Dogs. Proceedings of the National Academy of Sciences, 116(29), 14677-14681. https://doi.org/10.1073/pnas.1820653116;

Labuguen, R., Bardeloza, D. K., Negrete, S. B., Matsumoto, J., Inoue, K., Shibata, T. (2019). Primate Markerless Pose Estimation and Movement Analysis Using DeepLabCut. In Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 297-300. https://doi.org/10.1109/ICIEV.2019.8858533;

Labuguen, R., Matsumoto, J., Negrete, S. B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K., Shibata, T. (2021). MacaquePose: a Novel "in the Wild" Macaque Monkey Pose Dataset for Markerless Motion Capture. Frontier in Behavioral Neuroscience, 14, ISSN: 1662-5153. https://doi.org/10.3389/fnbeh.2020.581154;

Lapp, H. E., Salazar, M. G., Champagne, F. A. (2023). Automated maternal behavior during early life in rodents (AMBER9 pipeline. Scientific Reports, 2023(13). https://doi.org/10.1038/s41598-023-45495-4;

Lencioni, G. C., de Sousa, R. V., de Souza Sardinha, E. J., Corrêa, R. R., Zanella, A. J. (2021). Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling. PloS one, 16(10), e0258672. https://doi.org/10.1371/journal.pone.0258672

Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J. (1998). Coding Facial Expressions with Gabor Wavelets. In Proceedings Third IEEE international conference on automatic face and gesture recognition, 200-205. https://doi.org/10.1109/AFGR.1998.670949;

Mahmoud, M., Lu, Y., Hou, X., McLennan, K., Robinson, P. (2018). Estimation of pain in sheep using computer vision. In: Moore, J. R., Handbook of Pain and Palliative Care: Biopsychosocial and Environmental Approaches for the Life Course, 145-157.

Maretti, G., Sorrentino, V., Finomana, A., Gamba, M., & Giacoma, C. (2010). Not just a pretty song: an overview of the vocal repertoire of *Indri indri*. Journal of Anthropological Sciences, 88, 151-165. PMID: 20834055

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., Bethge, M. (2018). DeepLabCut: Markerless Pose Estimation of User-Defined Body Parts with Deep Learning. Nature Neuroscience, 21, 1281-1289. https://doi.org/10.1038/s41593-018-0209-y;

McKnight, P. E., Najab, J. (2010). Mann-Whitney U Test. The Corsini Encyclopedia of Psychology. https://doi.org/10.1002/9780470479216.corpsy0524;

Mootnick, A. R., Fan, P. F. (2011). A comparative study of crested gibbons (Nomascus). *American Journal of Primatology*, *73*(2), 135-154. https://doi.org/10.1002/ajp.20880

Morozov, A., Parr, L. A., Gothard, K., Paz, R., Pryluk, R. (2021). Automatic recognition of macaque facial expressions for detection of affective states. eneuro, 8(6). https://doi.org/10.1523/ENEURO.0117-21.2021

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., Mathis, M. W. (2019). Using DeepLabCut for 3D Markerless Pose Estimation Across Species and Behaviors. Nature Protocols, 14, 2152-2176. https://doi.org/10.1038/s41596-019-0176-0;

Ning, J., Li, Z., Zhang, X., Wang, J., Chen, D., Liu, Q., Sun, Y. (2022). Behavioral Signatures of Structured Feature Detection During Courtship in *Drosophila*. Current Biology, 32(6), 1211-1231. https://doi.org/10.1016/j.cub.2022.01.024;

Parr, L. A., Waller, B. M., Vick, S. J., Bard, K. A. (2007). Classifying chimpanzee facial expressions using muscle action. Emotion, 7(1), 172. https://doi.org/10.1037/1528-3542.7.1.172;

Paulet, J., Molina, A., Beltzung, B., Suzumura, T., Yamamoto, S., Suer, C. (2024). Deep learning for automatic facial detection and recognition in Japanese macaques: illuminating social networks. Primates: 65, 265–279. https://doi.org/10.1007/s10329-024-01137-5

Popelinsky, L., Brazdil, P. (2000). The principal components method as a pre-processing stage for decision tree learning, in: Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, Lyon, France;

Popik, P., Cyrano E., Piotrowska, D., Holuj M., Golebiowska, J., Malikowska-Racia, N., Potasiewicz, A., Nikiforuk, A. (2023). Effects of ketamine on rat social behavior as analyzed by DeepLabCut and SimBA deep learning algorithms. Frontiers in Pharmacology, 2023(14). https://doi.org/10.3389/fphar.2023.1329424;

Ramachandra, T. V., Mondal, T., Setturu, B. (2023). Relative performance evaluation of machine learning algorithms for land use classification using multispectral moderate resolution data. SN Applied Sciences, 5(10), 274. https://doi.org/10.1007/s42452-023-05496-4

Ravaglia, D., Ferrario, V., De Gregorio, C., Carugati, F., Raimondi, T., Cristiano, W., Torti, V., von Hardenberg, A., Ratsimbazafy, J., Valente, D., Giacoma, C., Gamba, M. (2023). There You Are! Automated Detection of Indris' Songs on Features Extracted from Passive Acoustic Recordings. Animals, 13(2):241. https://doi.org/10.3390/ani13020241;

Roßbach, P. (2018). Neural networks vs. random forests–does it always have to be deep learning. Germany: Frankfurt School of Finance and Management.

Sato, Y., Kondo, T., Uchida, A., Sato, K., Yoshino-Saito, K., Nakamura, M., Okano, H., Ushiba, J. (2022). Preserved intersegmental coordination during locomotion after cervical spinal cord injury in common marmosets. Behavioural Brain Research, 425:113816. https://doi.org/10.1016/j.bbr.2022.113816;

Schofield, D. P., Albery, G. F., Firth, J. A., Mielke, A., Hayashi, M., Matsuzawa, T., Carvalho, S. (2023). Automated Face Recognition Using Deep Neural Networks Produces Robust Primate Social Networks and Sociality Measures. Methods in Ecology and Evolution, 14(8), 1937-1951 https://doi.org/10.1111/2041-210X.14181;

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., Carvalho, S. (2019). Chimpanzee Face Recognition from Videos in the Wild Using Deep Learning. Science Advances, 5(9): eaaw0736. https://doi.org/10.1126/sciadv.aaw0736;

Shapiro, S. S., Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591-611. https://doi.org/10.2307/2333709;

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F,. Ghamisi, P., Homayouni, S. (2020) Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13,. 6308-6325, doi: 10.1109/JSTARS.2020.3026724

Shrout, P. E., Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, 86(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420;

Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J. T., Burton, C., Townsend, S. E., Carbone, C., Rowcliffe, J. M., Whittington, J., Brodie, J., Royle, J. A., Switalski, A., Clevenger, A. P., Heim, N., Rich, L. N. (2017). Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. Frontiers in Ecology and the Environment, 15(1), 26-34. https://doi.org/10.1002/fee.1448

Štrumbelj, E., Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41, 647-665. https://doi.org/10.1007/s10115-013-0679-x

Student. (1908). The probable error of a mean. Biometrika, 1-25. https://doi.org/10.2307/2331554

Sugai, L. S. M., Silva, T. S. F., Ribeiro, J. W., Llusia, D. (2018). Terrestrial passive acoustic monitoring: review and perspectives. BioScience, 69, 15–25. https://doi.org/10.1093/biosci/biy147

Suryanto, M. E., Luong, C. T., Vasquez, R. D., Roldan M. J. M., Hung, C. H., Ger, T. R., Hsiao C. D. (2023). Using crayfish behavior assay as a simple and sensitive model to evaluate potential adverse effects of water pollution: Emphasis on antidepressants. Ecotoxicology and Environmental Safety, 2023(265), ISSN: 0147-6513. https://doi.org/10.1016/j.ecoenv.2023.115507;

Taglialatela, J. P., Russell, J. L., Pope, S. M., Morton, T., Bogart, S., Reamer, L. A., Schapiro, S. J., Hopkins, W. D. (2015). Multimodal communication in chimpanzees. American journal of primatology, 77(11), 1143-1148. https://doi.org/10.1002/ajp.22449

Tirelli, T., Favaro, L., Gamba, M., & Pessani, D. (2011). Performance comparison among multivariate and data mining approaches to model presence/absence of Austropotamobius pallipes complex in Piedmont (North Western Italy). Comptes rendus biologies, 334(10), 695-704. https://doi.org/10.1016/j.crvi.2011.07.002

Tirelli, T., & Pessani, D. (2011). Importance of feature selection in decision-tree and artificial-neural-network ecological applications. *Alburnus alburnus alborella*: A practical example. Ecological informatics, 6(5), 309-315. https://doi.org/10.1016/j.ecoinf.2010.11.001

Tomar, S. (2006). Converting Video Formats with FFmpeg. Linux Journal, 2006(146), ISSN: 1075-3583. LinuxJ;

Tseng, S. P., Hsu S. E., Wang J. F., Jen I-F. (2024). An Integrated Framework with ADD-LSTM and DeepLabCut for Dolphin Behavior Classification. Journal of Marine Science and Engineering, 12(4), 540. https://doi.org/10.3390/jmse12040540**;**

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M.C., Stewart, C. V., Berger-Wolf, T., (2022). Perspectives in machine learning for wildlife conservation. Nature Communication, 13, 792 https://doi.org/10.1038/s41467-022-27980-y

Tuttle, A. H., Molinaro, M. J., Jethwa, J. F., Sotocinal, S. G., Prieto, J. C., Styner, M. A., Mogil, J. S., Zylka, M. J. (2018). A deep neural network to assess spontaneous pain from mouse facial expressions. Molecular pain, 14. https://doi.org/10.1177/1744806918763658

Uddin, S., Khan, A., Hossain, M.E., Moni, M.A.. (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19, 281 . https://doi.org/10.1186/s12911-019-1004-8

Valente, D., Miaretsoa, L., Anania, A., Costa, F., Mascaro, A., Raimondi, T., De Gregorio, C., Torti, V., Friard, O., Ratsimbazafy, J., Giacoma, C., Gamba, M. (2022). Comparative analysis of the vocal repertoires of the indri (Indri indri) and the diademed sifaka (Propithecus diadema). International Journal of Primatology, 43(4), 733-751.

Vick, S. J., Parr, B. M. (2007). Cross-species Comparison of Facial Morphology and Movement in Humans and Chimpanzees Using the Facial Action Coding System (FACS). Journal of Nonverbal Behavior, 31, 1-20. https://doi.org/10.1007/s10919-006-0017-z;

Waller, B. M., Julle-Daniere, E., Micheletta, J. (2020). Measuring the Evolution of Facial 'Expression' Using Multi-Species FACS. Neuroscience and Biobehavioral Reviews, 113, 1-11. https://doi.org/10.1016/j.neubiorev.2020.02.031;

Waller, B. M., Kavanagh, E., Micheletta, J., Clark, P. R., Whitehouse, J. (2022). The face is central to primate multicomponent signals. International Journal of Primatology, ISSN: 0164-0291. https://doi.org/10.1007/s10764-021-00260-0;

Waller, B. M., Parr, L. A., Gothard, K. M., Burrows, A. M., Fuglevand, A. J. (2008). Mapping the contribution of single muscles to facial movements in the rhesus macaque. Physiology and Behavior, 95(1-2), 93-100. https://doi.org/10.1016/j.physbeh.2008.05.002;

Wei, X., Wang, H., Scotney. B., Wan, H. (2020). Minimum margin loss for deep face recognition. Pattern Recognition. 2020(97). https://doi.org/10.1016/j.patcog.2019.107012

Wiltshire, C., Lewis-Cheetham, J., Komedová, V., Matsuzawa, T., Graham, K. E., Hobaiter, C. (2023). DeepWild: Application of the Pose Estimation Tool DeepLabCut for Behaviour Tracking in Wild Chimpanzees and Bonobos. Journal of Animal Ecology, 92(8), 1560-1574. https://doi.org/10.1111/1365-2656.13932;

Wilke, C., Kavanagh, E., Donnellan, E., Waller, B. M., Machanda, Z. P., Slocombe, K. E. (2017). Production of and responses to unimodal and multimodal signals in wild chimpanzees, *Pan troglodytes schweinfurthii*. Animal Behaviour, 123, 305-316. https://doi.org/10.1016/j.anbehav.2016.10.024

Witham, C. L. (2018). Automated Face Recognition of Rhesus Macaques. Journal of Neuroscience Methods, 300, 157-165. https://doi.org/10.1016/j.jneumeth.2017.07.020;

Wrench, A., Balch-Tomes, J. (2022). Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut. Sensors, 22(3). https://doi.org/10.3390/s22031133;

Yehia, H. C., Kuratate, T., Vatikiotis-Bateson, E. (2002). Linking Facial Animation, Head Motion and Speech Acoustics. Journal of Phonetics, 30(3), 555-568. https://doi.org/10.1006/jpho.2002.0165;

Yehia, H. C., Rubin, P., Vatikiotis-Bateson, E. (1998). Quantitative Association of Vocal-Tract and Facial Behavior. Speech Communication, 26(1/2), 23-43. https://doi.org/10.1016/S0167-6393(98)00048-X;

Zhang, K., Zhang, Z., Li, Z., Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights should be submitted in a separate editable file in the online submission system. Please use 'Highlights' in the file name and include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).

- DeepLabCut represents a promising tool for quantifying facial movements.
- We can automatically discriminate among voiced and unvoiced primate faces.