Same data, different results? Evaluating machine learning approaches for individual identification in animal vocalisations

K Wierucka[1,2*], D Murphy[3,4], SK Watson[5], N Falk[6,7], C Fichtel[1], J León[8], ST Leu[9], PM Kappeler[1], EF Briefer[10], MB Manser[6,7], N Phaniraj[7,11], M Scheumann[12], JM Burkart[2,6]

[1]German Primate Center - Leibniz Institute for Primate Research, Behavioral Ecology and Sociobiology Unit, Kellnerweg 4, 37077 Göttingen, Germany

[2]Department of Evolutionary Anthropology, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[3]German Primate Center - Leibniz Institute for Primate Research, Cognitive Ethology Laboratory, Kellnerweg 4, 37077 Göttingen, Germany

[4]Georg-August-Universität Göttingen, Johann-Friedrich-Blumenbach Institute, Department for Primate Cognition, Kellnerweg 4, 37077 Göttingen, Germany

[5]Department of Comparative Language Science, University of Zurich, Affolternstrasse 56, 8050 Zürich, Switzerland

[6]Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Affolternstrasse 56, 8050 Zürich, Switzerland

[7]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[8]Max Planck Institute of Animal Behaviour, Department for the Ecology of Animal Societies, Konstanz, Germany

[9]School of Animal and Veterinary Sciences, The University of Adelaide, Mudla Wirra Road, Roseworthy 5371, Australia

[10]Behavioural Ecology Group, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark

[11]Neuroscience Center Zurich (ZNZ), University of Zurich and ETH Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[12]University of Veterinary Medicine Hannover, Institute of Zoology, Bünteweg 17, 30559 Hannover

*Corresponding author: kwierucka@dpz.eu

ABSTRACT

Automated acoustic analysis is increasingly used in animal communication studies, and determining caller identity is a key element for many investigations. However, variability in feature extraction and classification methods limits the comparability of results across species and studies, constraining conclusions we can draw about the ecology and evolution of the groups under study. We investigated the impact of using different feature extraction (spectro-temporal measurements, Mel-frequency cepstral coefficients, and highly comparative time-series analysis) and classification methods (discriminant function analysis, support vector machines, Gaussian mixture models, neural networks, and random forests) on the consistency of classification accuracy across 16 mammalian datasets. We found that Mel-frequency cepstral coefficients and random forests yield consistently reliable results across datasets, facilitating a standardised approach across species that generates directly comparable data. These findings remained consistent across vocalisation sample sizes and number of individuals considered. We offer guidelines for processing and analysing mammalian vocalisations, fostering greater comparability, and advancing our understanding of the evolutionary significance of acoustic communication in diverse mammalian species.

Keywords: machine learning, bioacoustics, individual identification, call distinctiveness, vocal communication, review, method comparison.

INTRODUCTION

Many animals produce individually distinct calls – vocalizations that contain unique characteristics or features, allowing for differentiation and identification of individuals[1]. Conveying information about individual identity is widespread among taxa, and particularly prevalent in social species, as identity plays a pivotal role in regulating various social interactions[2,3]. The ability for researchers to accurately and reliably identify individuals based on their acoustic cues is also immensely important, serving as a gateway to a myriad of applications, e.g., to gain insights into the behaviour and ecology of species or to inform conservation and management strategies[4–6].

Quantifying the degree of individual identity information that is encoded in a particular vocalisation is difficult. Broadly, the process requires recording sound (multiple vocalisations per individual on separate occasions), extracting and quantifying their acoustic features ('feature extraction') from the recording and using these features to classify the vocalisation as belonging to one of a number of possible individuals[7]. Recently, with greater computational power becoming more and more accessible, new, automated methods for feature extraction and classification are becoming increasingly popular[8–10]. The ease of application of various classifiers, including those using machine learning techniques, has provided many opportunities within the field of bioacoustics[11–13]. As a consequence, a plethora of methods is currently used with no consensus or consistent protocol[14,15] (although, some general guidelines and good practices have been put forward[7,15,16]). One reason could be that individuality can be encoded in different ways, which varies by species and context[12], making it difficult to generalise methods[12]. Furthermore, the set of methods that can be used is influenced by various limitations of each particular dataset and the research questions to be addressed. Consequently, most researchers need to develop their own pipelines from

scratch, which leads to various, non-compatible and often unjustified methods for investigating individual identity[7]. This makes it nearly impossible make comparisons across studies or expand conclusions beyond one study system.

To identify the diversity of approaches in use, we systematically reviewed research articles focusing on empirical investigations of mammalian species incorporating acoustic analysis and evaluating the individual distinctiveness of vocalisations (details in Supplementary Materials (SM), section 2). We obtained a total of 119 observation points and found a considerable diversity in feature extraction methods used across studies (Figure 1, SM section 2). While many studies relied on spectro-temporal features to ascertain the presence of individual distinctiveness, the specific measurements employed varied greatly, with 41 unique parameter sets used within this feature extraction method across studies (SM, dataset 1). We found that 24.37% of studies provided external (i.e., citation-based) justification for their approach towards methods used, 27.73% described it as being based on their prior experience with the system (internal justification), and 47.9% provided no justification at all (Figure 1, SM section 2).

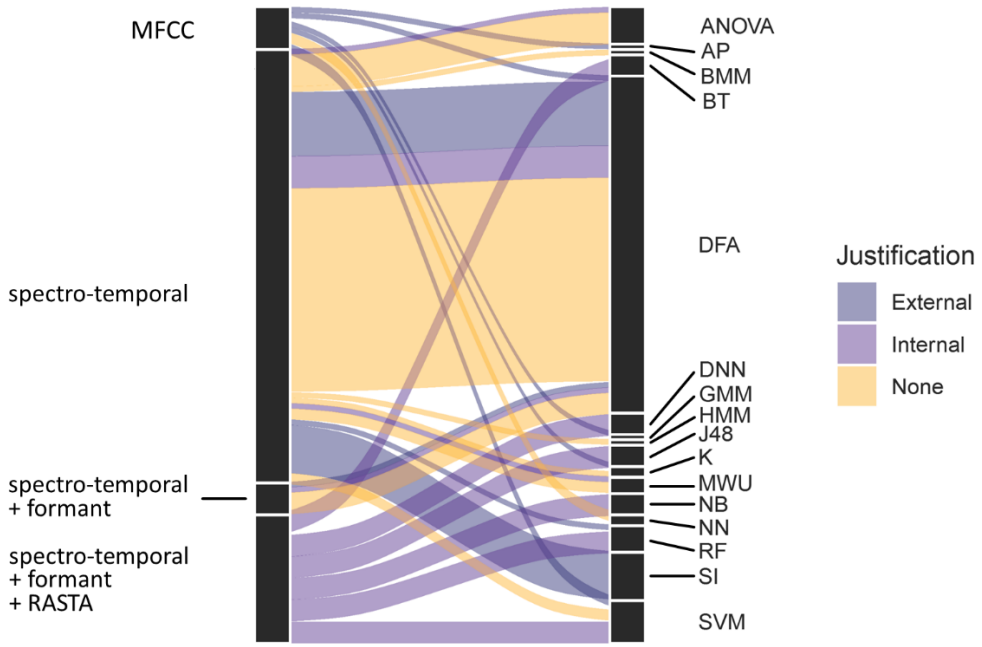

Figure 1. Acoustic feature extraction methods and classifiers used to evaluate individual distinctiveness in mammalian vocalisations. Notations: ANOVA – analysis of variance – traditional or multivariate; AP – affinity propagation; BMM – Bayesian functional mixed model; BT – bagged trees; DFA – discriminant function analysis (normal and permutational); DNN – deep neural network; DTW – dynamic time warping; GMM – Gaussian mixture model; HMM – hidden Markov model; K – kmeans or kmedioids; MWU – Mann-Whitney-U test; NB – naïve Bayes; NN - neural network; RF – random forest; SI – stereotypy Index; SVM – support vector machine; MFCC – Mel frequency cepstral coefficients; spectro-temporal – frequency and temporal parameters extracted from the spectrogram, RASTA – relative spectral processing, External justification – based on published work; Internal justification – based on prior experience with the study system. N articles = 46, N datapoints = 119 (as certain studies encompassed multiple analyses, call types, or species). See SM section 2 for details.

There is currently no published information about the potential variation in results that may derive from using different extraction and classification method combinations for determining individual identity in bioacoustics. Previous research has drawn attention to this issue and call for a unification of the methods used[16,17] and several excellent studies have investigated the impact of different statistics[7,14] or feature extraction methods[18] on the degree of individual identity quantified in vocalizations[18]. However, none of the above have considered the combined effects of feature extraction and classification methods on results. A comparison of the consistency of results obtained using various extraction and classification methods, and a validation of the relationship between these would provide a first step towards the unification of methodologies on vocal individuality, thus ensuring robust and reliable, comparable inferences about the data and, by extension, animal communication, and connected fields.

In this paper, using data spanning a diverse range of mammalian taxa, we undertook a broadscale analysis to assess the relative differences in results obtained through various combinations of feature extraction and classification methods. Additionally, we examined the influence of sample size (number of vocalisations per individual) and number of classes (number of individuals to classify) on these outcomes. The main aim was to establish which methods consistently produce the least amount of variation in results, in order to identify the feature extraction and classification methods that produce robust results. Finally, we presented guidelines for future studies, with the goal of fostering a more unified approach to the investigation of animal bioacoustics.

METHODS

Data collection

To make our findings as widely applicable as possible, we used data from 14 mammalian species: alpaca (*Vicugna pacos*), domestic cat (*Felis catus*), Egyptian fruit bat (*Rousettus aegyptiacus*), domestic goat (*Capra aegagrus hircus*), domestic horse (*Equus caballus*), sooty mangabey (*Cercocebus atys*), common marmoset (*Callithrix jacchus*), meerkat (*Suricata suricatta*), grey mouse lemur (*Microcebus murinus*), domestic pig (*Sus scrofa domesticus*), redfronted lemur (*Eulemur rufifrons*), domestic sheep (*Ovis aries*), squirrel monkeys (*Saimiri sciureus*) and tree shrew (*Tupaia belangeri*) (details of the data collection and processing for each species are available in the SM section 3). For two species (marmosets and mouse lemurs), we had recordings of two distinct call types, and these were treated as independent datasets. Consequently, our analysis incorporated a total of 16 unique datasets, which we refer to as 'species' in the model descriptions. For each species, data were limited to the same call type (all were of relatively constant frequency) and each file contained a single call (call bouts were not included in the analysis), with the recording cropped tightly around the vocalisation. Caller identity was known for all calls. Datasets varied in the number of calls per individual and the number of individuals they included, thus, we prepared two separate datasets for the different analyses we conducted. The first was used for analyses investigating differences in results given different sample sizes and number of individuals and consisted of 4 mammalian species representing diverse taxa: Egyptian fruit bat, common marmoset (phees), meerkat, and domestic sheep. We included vocalisations of 9 individuals per species and 60 calls per individual (all randomly selected; dataset 1, Figure 2). Such a large number of recordings per individual is often difficult to obtain, and we wanted to include a wider range of species in our analysis**.** Thus, the second subset included data from all species, however we randomly selected 9 individuals and 10 calls per individual (dataset 2, Figure 2).

Feature extraction

The source and filter theory[19,20] links vocal production to the acoustic structure of cues. In mammals, the glottal wave is produced in the larynx (source) through the vibration of vocal folds and is then modified by the vocal tract (filter)[20]. Therefore, source-related features typically refer to measurements derived from the fundamental frequency, duration and amplitude contour (measured from the spectrogram), while the filter-related parameters include those related to the spectral envelope of the acoustic signal[21], measured as formant frequencies or Mel-frequency cepstral coefficients (MFCC). Since individual identity can be encoded both in source- and filter-related measurements [21], we include both categories of features in our analysis.

Spectrogram features were extracted for all datasets using Raven Pro 1.6.4[22]. To limit bias resulting from different researchers processing data, we extracted robust measurements from the calls (Centre Frequency, Frequency 5%, Frequency 25%, Frequency 75%, Frequency 95%, Duration 90%, Duration 50%, Bandwidth 50%, Bandwidth 90%[23] which are based on the energy stored within the sound rather than boundaries of a selection (and are therefore less sensitive to user variability), as well as average and aggregate entropy measures (spectro-temporal dataset). We extracted MFCC for all datasets (package BehaviouR[24]) from the recorded vocalisations (frequency minima and maxima defined separately for a for each species; SM, Table S2) (MFCC dataset). In addition, we extracted features using a new approach – the Highly Comparative Time Series Analysis (HCTSA)[25,26], which extracts thousands of time-series features directly from the wave form (HCTSA dataset). Consequently, three feature extraction methods (spectro-temporal, MFCC and HCTSA; Figure 2) were used and included in further analysis.

Classification

We applied four commonly used classifiers (discriminant function analysis (DFA), neural networks (NN), support vector machines (SVM), random forests (RF)) to each dataset (spectro-temporal, MFCC and HCTSA) for each species, to determine how accurately each one assigns calls to individuals (Figure 2). As the goal was not to obtain the most accurate model possible, but to determine how comparably they performed, we did not optimise the models to each dataset separately. For all analyses, data were split 70:30 into training and testing sets (within individual and with no overlap), with results reported for testing sets only.

For SVMs, we implemented a radial basis kernel type, using a modified version of the trainSVM function (GibbonR package[27]; the modification split the data without overlap). This function automatically tunes gamma and cost parameters. We ran a RF model (randomForest::tuneRF[28]) using the optimal mtry found (doBest=TRUE), with 1000 trees used at the tuning step. Neural networks can be affected by collinearity, therefore, we performed a PCA (stats::prcomp; with variables zero-centered and scaled to have unit variance) prior to running a Model Averaged Neural Network (modified version of the trainNN function; modification as above; GibbonR package[27]). Before classifying the data using DFA (MASS::lda[29]), we also conducted a PCA (as above) and the data were normalised (bestNormalize::bestNormalize[30]). The DFA has many assumptions, which we checked against our data: Our groups were independent and mutually exclusive, and sample size was balanced. Potential collinearity issues were resolved by applying DFA only to data following a PCA. Multivariate normality (MVN::mvn[31]) and homogeneity of covariances (biotools::boxM[32]) were tested. While we tested for the assumptions of a DFA, in order to gauge the

robustness of the existing literature's findings, we extended our analysis to include datasets that did not strictly satisfy the DFA assumptions.

Based on the confusion matrices obtained from running the classifiers, we calculated two widely used evaluation metrics: accuracy and F1 scores. Accuracy gives a measure of the proportion of correctly classified samples returned by a classifier, but in some situations can be misleading with regard to the amount of incorrectly classified samples, particularly with imbalanced classes. F1 scores are often more useful in "real world" situations, where classes are imbalanced and there is a high cost associated with incorrect classifications. Here, we have a set of multi-class classification problems with balanced classes (i.e., the same number of calls per individual), so we use accuracy computed as the sum of the diagonal elements of the normalised confusion matrix as this accounts for total number of possible classes (i.e., individuals). We used caret::confusionMatrix[33] to compute both accuracy and F1 scores. To account for variation in the results due to initial conditions of the classifiers, we ran 20 iterations (20 seeds) of training for each model, each with initial conditions set under a different random seed value.

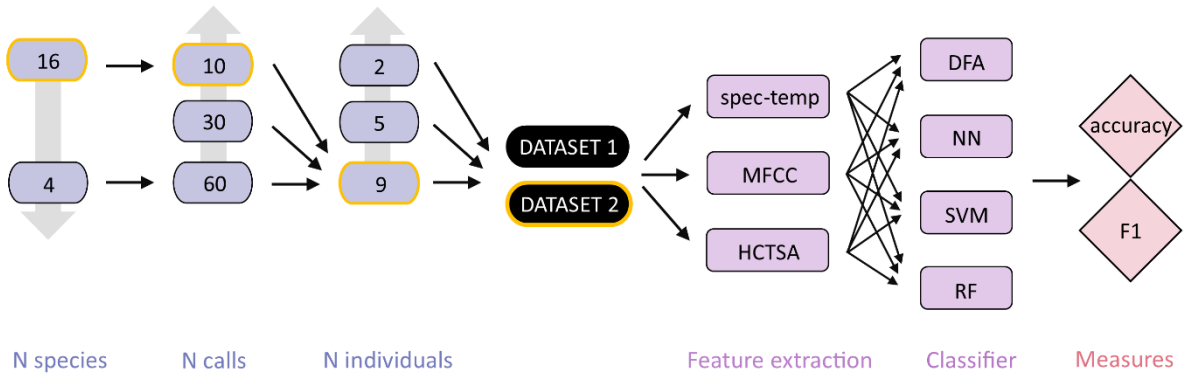

Figure 2. Sequential steps of the methodology used for evaluating the robustness of individual identity estimation in mammal vocalisations. Notations: spec-temp – spectro-temporal, MFCC – Mel-frequency Cepstral Coefficients, HCTSA - Highly Comparative Time Series Analysis, DFA – Discriminant Function Analysis, NN – Neural Networks, SVM – Support Vector Machines, RF – Random Forest. Grey arrows indicate a subset of the former group. Orange outlines indicate data used in dataset 2.

Statistical analysis

All analyses were conducted in R version 4.2.2[34], with linear models fitted using glmmTMB:glmmTMB[35]. All P-values we reported are based on likelihood ratio tests (LRT)[36]. For each analysis, we fitted a GLMM with beta error distribution and logit link function. Prior to fitting each model, we transformed the response to avoid values being exactly zero or one[37].

In order to avoid overconfident model estimates and to keep type I error rate at the nominal level of 0.05, we included random slopes[38,39] of the fixed effects within each of the random factors in each model. For all models, we originally included parameters for the correlations among random slopes and intercepts, but as the models did not converge, we had to exclude them. To include the fixed effect terms as random slopes, we therefore manually dummy coded and centred them to a mean of zero. See below for specific details of each model. Most of the models also did not converge with the interactions between the fixed

effect terms as random slopes, so we also excluded the interaction terms from the random slopes, except where noted below.

We determined model stability of each model by dropping individual levels of each random factor from the data, one at a time, and comparing the estimates derived from models fitted on the respective subsets with those obtained for the full dataset and found the models to be of acceptable stability (see results).

Consistency of classification performance

Our first aim was to estimate how classification accuracy and F1 scores vary depending on the choice of extraction method and classifier. Using a dataset of 4 species, with 9 individuals per species and 60 calls per individual, we fitted one GLMM, with accuracy as the response variable and another with F1 as the response variable. In both models, we included classifier (a factor with four levels), extraction method (a factor with three levels) and their interaction as fixed effects and a crossed random effect structure with one random intercept of species (a factor with 4 levels) and another of seed (factor with 20 levels). In both models, we were able to include the interaction between the fixed effect terms in the random slopes.

As an overall test of the effect of classifier and its interaction with extraction method, and to avoid 'cryptic multiple testing'[40], we compared each model with an appropriate null model lacking the classifier term in the fixed effects but being otherwise identical. We obtained confidence intervals (CIs) of model estimates and fitted values by means of a parametric bootstrap (N=1,000 bootstraps; function simulate of the package glmmTMB; note that for the accuracy model, only 717 bootstrapped models successfully converged and for the F1 model, only 753 converged. The models were not overdispersed (accuracy: dispersion parameter = 0.975; F1: dispersion parameter = 0.904). The samples analysed with these model comprised a total of 960 scores for each evaluation metric (i.e., 4 call types x 3 extraction methods x 4 classifiers x 20 iterations).

Effect of sample size of vocalisations: n60 vs n30 vs n10

As high-quality data collected from animals can often be limited in sample size, we also examined how the observed accuracy of the classifiers varied with the sample size (i.e., number of calls per individual) included. We created subsets of the data (samples selected randomly, where a different random subset was selected for each seed value) that varied in sample size, with 60 (n60 dataset), 30 (n30 dataset) and 10 (n10 dataset) calls per individual (the number of individuals remained the same at n=9 per species). In the model, we included classifier (a factor with four levels), extraction method (a factor with three levels), sample size (a factor with three levels) and the 3-way interaction between them as a fixed effect. We used a crossed random effect structure, with one random intercept effect of species (a factor with four levels) and another of seed (a factor with 20 levels).

As an overall test of the effect of sample size and its interaction with extraction method and classifier, and to avoid 'cryptic multiple testing'[40], we compared the model with an appropriate null model lacking the sample size term in the fixed effects but being otherwise identical.

We obtained CIs of model estimates and fitted values by means of a parametric bootstrap using glmmTMB::simulate (N=1000 bootstraps; note that only 855 of the bootstrapped models converged). The

model was not overdispersed (dispersion parameter=1.02). The samples analysed with this model comprised a total of 2880 accuracy scores (i.e., 4 species x 3 extraction methods x 4 classifiers x 3 sample size sets x 20 iterations).

Effect of number of individuals – 9 vs 5 vs 2 individuals

To determine the consistency of classifier performance when there are fewer possible classes to assign, we ran a similar analysis to the one described above, using subsets of the data restricted in this case by the number of individuals rather than the number of calls per individual. In the first subset, we randomly selected 5 individuals from each of the 4 species, and in the second, we randomly selected 2 individuals. This resulted in 3 datasets per seed, with 9, 5 and 2 individuals per species respectively. Each dataset contained 60 calls per individual, and in each of the 20 iterations, a different set of individuals was randomly selected.

As an overall test of the effect of number of individuals and its interaction with extraction method and classifier, we compared the model with an appropriate null model lacking the number of individuals term in the fixed effects but being otherwise identical.

CIs of model estimates were calculated as above (953 of the bootstrapped models for the successfully converged). The model was not overdispersed (dispersion parameters =1.02). The samples analysed with this model comprised a total of 2880 accuracy scores (i.e., 4 species x 3 extraction methods x 4 classifiers x 3 sample size training sets x 20 iterations).

Consistency of effects on larger taxonomic range

Using data from all 16 species, we repeated the analysis described in the *Consistency of classification performance* section above. In the accuracy model, we were able to include the interaction between the fixed effect terms in the random slopes. 908 and 792 of the bootstrapped models successfully converged for the accuracy and F1 models, respectively. The models were not overdispersed (model dispersion parameters: accuracy model = 0.92; F1 model = 0.91). The samples analysed with these models each comprised a total of 3840 scores for each metric (i.e., 16 species x 3 extraction methods x 4 classifiers x 20 iterations).

RESULTS

Consistency of classification performance

Overall, the full-null model comparisons revealed a clear effect of classifier in each model (accuracy $\chi2=61.917$, df=9, $P<0.001$, F1 $\chi2=49.278$, df=9, $P<0.001$; SM, Table S3-S4). In each model, the interaction between classifier and extraction method was also significant (LRT: accuracy $\chi2=39.513$, df=6, $P<0.001$, F1 $\chi2=32.021$, df=6, $P<0.001$; SM, Table S3-S4)

For both evaluation metrics, model results revealed the same general pattern of performance for each combination of extraction method and classifier (Figure 3). Classifiers trained on data extracted using the

MFCC feature extraction method all typically provided the highest and most consistent scores (difference between maximum and minimum predicted values Δ = 0.047). A greater variation in the performance of classifiers was observed when using the spectro-temporal dataset (Δ = 0.184), and to an even greater extent, HCTSA feature extraction methods (Δ = 0.632). Out of classifiers, random forests resulted in the least amount of variation (Δ = 0.221) across various extraction methods, followed by SVM: Δ = 0.352). DFA and NN provided most inconsistent results, with DFA having the highest variation (Δ = 0.687; Figure 3; SM, Table S3-S4).

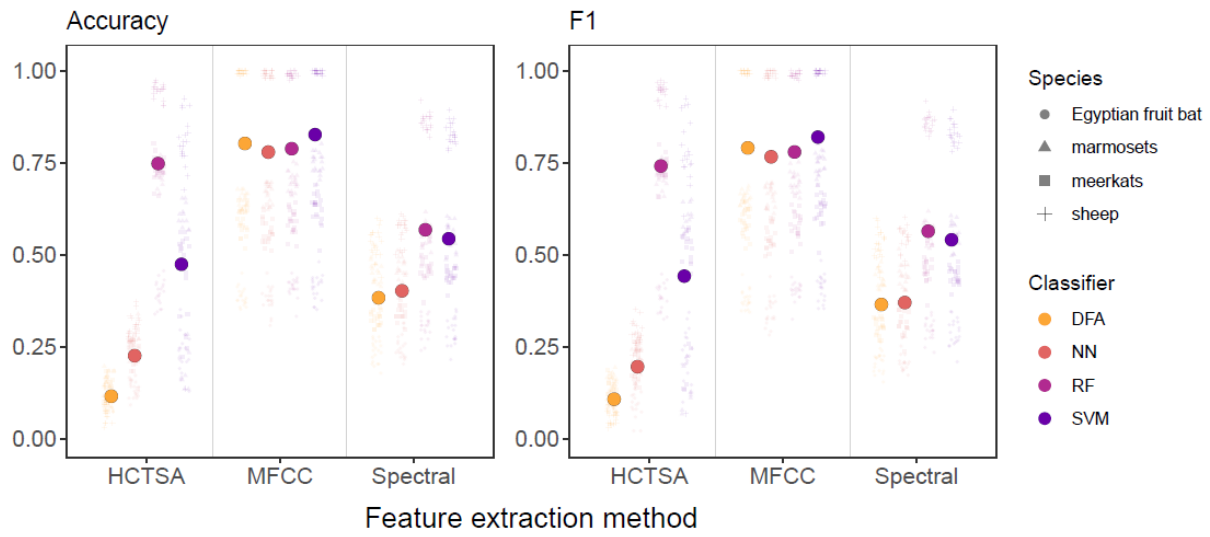

Figure 3. Observed (smaller points) and model predicted estimates (larger points) of accuracy and F1 scores for each classifier across various extraction methods for datasets with 4 species, 9 individuals per species and 60 calls per individual. Values were back transformed to the response scale.

<u>Effect of sample size of vocalisations on method performance – n60 vs n30 vs n10</u>

The full-null model comparison revealed a clear effect of sample size (number of vocalisations per individual) on accuracy ($\chi2$=1043.031, df=24, P<0.001) and the 3-way interaction between sample size, extraction method and classifier was significant (LRT: $\chi2$=581.121, df=12, P<0.001) (Figure 4A; SM, Table S5), with the same trends visible as described in the previous section (MFCC and RF providing most consistent results across sample sizes).

For most classifier-extraction method combinations, accuracy increased from n10 to n30, with not much of an increase from n30 to n60 (Figure 4A). The exception to this trend was the HCTSA extraction method paired with the DFA, where increasing the sample size of the training data set to n60 lead to a decrease in accuracy. MFCC datasets typically led to more consistent and better accuracy values than the other extraction methods for all sample sizes, regardless of which classifier they were paired with. Across classification methods, RF performed most consistently across extraction methods in datasets of all sample sizes (Figure 4A).

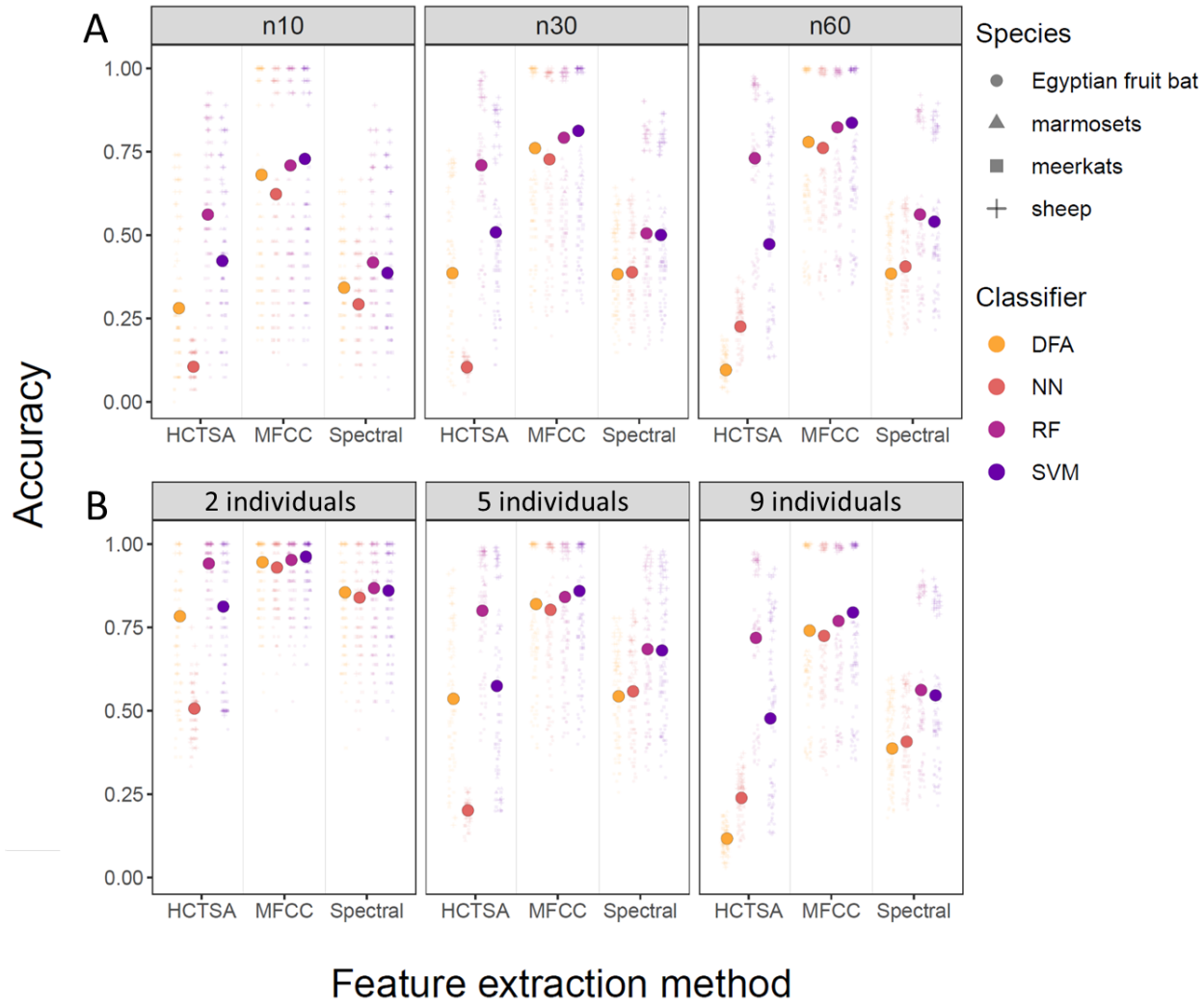

Figure 4. Observed (smaller points) and model predicted estimates (larger points) of accuracy scores for each classifier across three extraction methods for datasets of different A) sample sizes (n10, n30, n60 – sample size of 10, 30, or 60 calls per individual, respectively); B) number of classes (2, 5 and 9 individuals per species). Values were back transformed to the response scale.

<u>Effect of number of individuals – 9 vs 5 vs 2 classes</u>

The full-null model comparison revealed a clear effect of the number of individuals on accuracy ($\chi2$=525.281, df=24, P<0.001) and the LRT showed that the 3-way interaction between number of classes, extraction method and classifier was significant ($\chi2$=298.84, df=12, P<0.001). While the overall trends were similar to those reported above (MFCC and RF performed most consistently), we observed more variation in the absolute values of the accuracy achieved by the classifiers within and between sample size groups. Accuracy values were highest and more consistent when a smaller number of individuals was present (Figure 4B; SM, Table S6).

Consistency of effects on larger taxonomic range

The analysis of the expanded dataset comprising 16 species revealed trends consistent with the aforementioned results. For both models, the full-null model comparisons revealed a clear effect of classifier (accuracy: $\chi2=2309.115$, df=9, $P<0.001$; F1: $\chi2=266.323$, df=9, $P<0.001$) and the LRTs showed that the effects of the interactions between extraction method and classifier were significant (accuracy: $\chi2=2204.115$, df=9, $P<0.001$; F1: $\chi2=145.409$, df=6, $P<0.001$).

Although the predicted estimates of accuracy and F1 scores exhibited a slight decrease upon the inclusion of additional species in the analysis, the overall trends indicated that MFCC extraction method yielded the most consistent and highest values across all classifiers, while the RF classifier returned the most consistent values of all classifiers across extraction methods (SM, Figure S1, Table S7-S8).

DISCUSSION

We assessed the comparability of various combinations of feature extraction and classification methods commonly used for the identification of individual vocal signature, aiming to establish a guideline for future research endeavours that facilitate standardised inter-taxa comparisons and contribute to larger-scale analyses and overarching conclusions on vocal individuality. We investigated the accuracy of these methods in assigning identity across 16 call types from 14 different mammalian species. Our findings showed that using Mel-frequency cepstral coefficients to extract acoustic features, or random forests as classifiers provides relatively consistent results. We therefore recommend using these methods in future investigations to enhance comparability of results across species.

Obtained accuracy values varied significantly depending on the employed feature extraction methods and classifiers. Results obtained from datasets using MFCC (describing the spectral envelope of the sound) as the extraction method were most robust and provided the most consistent results across different classifiers used. As filter-related acoustic parameters (such as MFCC) of the vocalisations reflect individual morphology and cannot be easily modified without a flexible vocal tract, they tend to be a more stable and reliable cue and are therefore more predictable[21]. This information can be used by animals directly for individual recognition, or to make the recognition process more efficient by distinguishing broader categories of individuals first (e.g., size, sex, age) and then identifying the individual. Parameters related to the spectral envelope have been extensively used for voice recognition in humans and have been shown to be an important cue for recognition in other non-human primates[41–44] as well as other taxa[45].

Results obtained when using HCTSA datasets were the most variable. This can be attributed, at least in part, to the very large number of features returned by this extraction method. A challenge arises when limited sample sizes intersect with a large number of features, potentially impeding the algorithm's accuracy during classification (usually due to overfitting). Conversely, a lower number of features may offer less information per datapoint during classification, which may be the reason for the lower accuracy values from the spectro-temporal dataset. We deliberately refrained from constraining the number of features across datasets, recognising it as an intrinsic characteristic of the extraction method that warrants consideration in the choice of methodology used for each study. Additionally, it is crucial to acknowledge that a larger number of features facilitates a more detailed description of the sound, although it is difficult to say if all of these features are biologically relevant. The number of features to be

used for classification thus requires careful consideration on a case-by-case basis, contingent upon the specific objectives and constraints of each study.

Among the various classifiers, RF consistently yielded the least variable results across diverse datasets (SVM produced relatively high accuracy values, yet were more slightly variable then RF results across different extraction methods). This can be attributed to the algorithm's ease of use, requiring few assumptions of the data. Conversely, DFA (commonly employed in previous studies) and NN exhibited the lowest accuracy values. This outcome can likely be attributed to the stringent assumptions imposed by DFA, often unmet by the majority of datasets. Given the limited availability of studies providing information on assumption testing, there is a high likelihood that results derived from DFA may lack reliability, precluding meaningful comparisons with other studies and warranting caution in future analyses.

We investigated the effect of different sample sizes on the obtained accuracies. Overall trends of accuracy values remained consistent, even when reducing sample sizes to 10 calls per individual, albeit with slightly lower values compared to n30 and n60 datasets. Given the inherent challenges associated with acquiring acoustic data, particularly from wild animals, the optimal approach, if feasible, appears to be targeting approximately 30 calls per individual. However, it seems that there is no compelling need to further increase the sample size beyond this threshold. Importantly, a sample size of 10 calls per individual presents trends comparable to those obtained with larger sample sizes.

We varied the number of individuals, to be classified, and found that the main trends were similar to those described above, with MFCCs and RF having least variation in accuracy values. As expected, since the task of classifying vocalisations to fewer groups is easier, we observed greater and more consistent values for the smaller number of individuals, with the accuracy values for two individuals being notable greater than for five or nine individuals (with little difference between the latter two).

This synthesis provides a comprehensive overview of the current landscape of machine learning applications in individual identity classification of animal vocalisations. While recognising the importance of tailoring analyses to species-specific characteristics and the biological questions posed, our findings indicate that employing Mel-frequency cepstral coefficients and random forests enhances comparability across species. As an initial step in establishing guidelines for processing mammalian acoustic data, where possible, we recommend including these in analyses, to facilitate future comparative studies. We encourage authors to consider these analyses in conjunction with species- and question-specific approaches, contributing to the creation of a shared foundation. To support this approach, we provided the code for these analyses (SM, section 1). We hope that our contribution will facilitate a shift towards methodological consistency in the field of bioacoustics, arming future researchers with the knowledge to perform analyses that will produce the directly comparable outputs (both across species, and across studies within species) necessary to answer the most challenging research questions.

ACKNOWLEDGEMENTS

Funding: KW, SKW, NF, MBM, NP and JMB were partially or entirely funded by the NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888. CF thanks the SFB1528 “Cognition of interactions” for discussions.

AUTHOR CONTRIBUTIONS

KW – concept, data preparation, data analysis, data collection and consolidation, writing original draft, manuscript review
DM – data analysis, manuscript editing
NF – literature review scoring, data contributions, manuscript review
SW – sourcing data, manuscript editing
NP – literature review scoring, HCTSA processing, manuscript review
JMB, PK – manuscript review
CF, JL, STL, EB, MBM, MS – data contributions, manuscript review

REFERENCES

1. Yorzinski, J. L. The cognitive basis of individual recognition. *Current Opinion in Behavioral Sciences* vol. 16 53–57 Preprint at https://doi.org/10.1016/j.cobeha.2017.03.009 (2017).

2. Tibbetts, E. A. & Dale, J. Individual recognition: it is good to be different. *Trends in Ecology and Evolution* vol. 22 529–537 Preprint at https://doi.org/10.1016/j.tree.2007.09.001 (2007).

3. Wyman, M. T., Walkenhorst, B. & Manser, M. B. Selection levels on vocal individuality: strategic use or byproduct. *Current Opinion in Behavioral Sciences* vol. 46 Preprint at https://doi.org/10.1016/j.cobeha.2022.101140 (2022).

4. Gentry, K. E. *et al.* Bioacoustics in cognitive research: Applications, considerations, and recommendations. *Wiley Interdiscip Rev Cogn Sci* 11, 1–19 (2020).

5. Terry, A. M. R., Peake, T. M. & McGregor, P. K. The role of vocal individuality in conservation. *Frontiers in Zoology* vol. 2 Preprint at https://doi.org/10.1186/1742-9994-2-10 (2005).

6. Tuia, D. *et al.* Perspectives in machine learning for wildlife conservation. *Nature Communications* vol. 13 Preprint at https://doi.org/10.1038/s41467-022-27980-y (2022).

7. Linhart, P., Mahamoud-Issa, M., Stowell, D. & Blumstein, D. T. The potential for acoustic individual identification in mammals. *Mammalian Biology* vol. 102 667–683 Preprint at https://doi.org/10.1007/s42991-021-00222-2 (2022).

8. Mcloughlin, M. P., Stewart, R. & McElligott, A. G. Automated bioacoustics: Methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface* vol. 16 Preprint at https://doi.org/10.1098/rsif.2019.0225 (2019).

9. Mutanu, L., Gohil, J., Gupta, K., Wagio, P. & Kotonya, G. A review of automated bioacoustics and general acoustics classification research. *Sensors* 22, (2022).

10. Rutz, C. *et al.* Using machine learning to decode animal communication. *Science (1979)* 381, 152–155 (2023).

11. Valletta, J. J., Torney, C., Kings, M., Thornton, A. & Madden, J. Applications of machine learning in animal behaviour studies. *Animal Behaviour* vol. 124 203–220 Preprint at https://doi.org/10.1016/j.anbehav.2016.12.005 (2017).

12. Stowell, D., Petrusková, T., Šálek, M. & Linhart, P. Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions. *J R Soc Interface* 16, (2019).

13. Phaniraj, N., Wierucka, K., Zürcher, Y. & Burkart, J. M. Who is calling? Optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers. *J R Soc Interface* 20, (2023).

14. Linhart, P. *et al.* Measuring individual identity information in animal signals: Overview and performance of available identity metrics. *Methods Ecol Evol* 10, 1558–1570 (2019).

15. Odom, K. J. *et al.* Comparative bioacoustics: a roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews* 96, 1135–1159 (2021).

16. Oswald, J. N. *et al.* A Collection of Best Practices for the Collection and Analysis of Bioacoustic Data. *Applied Sciences (Switzerland)* vol. 12 Preprint at https://doi.org/10.3390/app122312046 (2022).

17. Keen, S. C. *et al.* A machine learning approach for classifying and quantifying acoustic diversity. 2021, 1213–1225 (2021).

18. Clemins, P. J., Johnson, M. T., Leong, K. M. & Savage, A. Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations. *J Acoust Soc Am* 117, 956–963 (2005).

19. Fant, G. *Acoustic Theory of Speech Production*. (Mouton & Co. N.V., Publishers, The Hague, 1970).

20. Titze, I. R. Vocal fold mass is not a useful quantity for describing F0 in vocalization. 54, 520–522 (2011).

21. Taylor, A. M., Charlton, B. D. & Reby, D. Vocal production by terrestrial mammals: source, filter, and function. in *Vertebrate sound production and acoustic communication* (eds. Suthers, R. A., Fitch, W. T., Fay, R. R. & Popper, A. N.) 229–259 (Springer International Publishing Switzerland, 2016). doi:10.1007/978-3-319-27721-9_8.

22. K. Lisa Yang Center for Conservation Bioacoustics. Raven Pro: Interactive Sound Analysis Software. Preprint at (2024).

23. Charif, R. A., Waack, A. M. & Strickman, L. M. Raven Pro 1.4 user's manual. *Cornell Lab of Ornithology, Ithaca, NY* 25506974, (2010).

24. Clink, D. behaviouR: An R package for online teaching of fundamental concepts in behavior and ecology. Preprint at (2023).

25. Fulcher, B. D., Little, M. A. & Jones, N. S. Highly comparative time-series analysis: The empirical structure of time series and their methods. *J R Soc Interface* 10, (2013).

26. Fulcher, B. D. & Jones, N. S. hctsa: a computational framework for automated time-Series phenotyping using massive feature extraction. *Cell Syst* 5, 527-531.e3 (2017).

27. Clink, D. gibbonR: An R package for the detection and classification of acoustic signals using machine learning. Preprint at (2023).

28. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* 2, 18–22 (2002).

29. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, New York, 2002).

30. Peterson, R. A. & Cavanaugh, J. E. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *J Appl Stat* 47, 2312–2327 (2020).

31. Korkmaz, S., Goksuluk, D. & Zararsiz, G. MVN: an R package for assessing multivariate normality. *R J* 6, 151–162 (2014).

32. da Silva, A. R., Malafaia, G. & de Menezes, I. P. P. biotools: an R function to predict spatial gene diversity via an individual-based approach. *Genetics and Molecular Research* 16, (2017).

33. Kuhn & Max. Building predictive models in R using the caret package. *J Stat Softw* 28, 1–26 (2008).

34. R Core Team. R: A Language and Environment for Statistical Computing. Preprint at https://www.r-project.org/ (2022).

35. Brooks, M. E. *et al.* ETH Library glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. doi:10.3929/ethz-b-000240890.

36. Dobson, A. J. *An Introduction to Generalized Linear Models*. (Chapman and Hall, London, UK, 2002).

37. Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods* (2006) doi:10.1037/1082-989x.11.1.54.supp.

38. Schielzeth, H. & Forstmeier, W. Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology* 20, 416–420 (2009).

39. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang* 68, 255–278 (2013).

40. Forstmeier, W. & Schielzeth, H. Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behav Ecol Sociobiol* 65, 47–55 (2011).

41. Gamba, M., Colombo, C. & Giacoma, C. Acoustic cues to caller identity in lemurs: a case study. *J Ethol* 30, 191–196 (2012).

42. Mielke, A. & Zuberbühler, K. A method for automated individual, species and call type recognition in free-ranging animals. *Anim Behav* 86, 475–482 (2013).

43. Clink, D. J., Crofoot, M. C. & Marshall, A. J. Application of a semi-automated vocal fingerprinting approach to monitor Bornean gibbon females in an experimentally fragmented landscape in Sabah, Malaysia. *Bioacoustics* 28, 193–209 (2019).

44. Clink, D. J., Tasirin, J. S. & Klinck, H. Vocal individuality and rhythm in male and female duet contributions of a nonhuman primate. *Curr Zool* 66, 173–186 (2020).

45. Townsend, S. W., Charlton, B. D. & Manser, M. B. Acoustic cues to identity and predator context in meerkat barks. *Anim Behav* 94, 143–149 (2014).