Remote Sensing in Ecology and Conservation
Open Access
ZSL

RESEARCH ARTICLE

A scalable transfer learning workflow for extracting biological and behavioural insights from forest elephant vocalizations

Alastair Pickering[1], Santiago Martinez Balvanera[1], Kate E. Jones[1] & Daniela Hedwig[2]

[1]Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK
[2]K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA

Keywords
African forest elephant, behavioural ecology, bioacoustics, passive acoustic monitoring, population ecology, transfer learning

Correspondence
Alastair Pickering, Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK. Tel.: +44 020 7679 2000; E-mail: alastair.pickering.22@ucl.ac.uk

Funding Information:
This research was supported by a Natural Environment Research Council grant to AP (grant number NE/S007229/1), a National Geographic Society (grant #28718 to DH) and a generous gift from K. Lisa Yang to the Cornell Lab of Ornithology. SMB was supported by a CONACYT grant (Grant/Award Number: 2020-000017-02EXTF-00334).

Editor: Dr. Vincent Lecours
Associate Editor: Dr. Emma Higgins

Received: 23 September 2024; Revised: 17 March 2025; Accepted: 18 March 2025

doi: 10.1002/rse2.70008

Abstract

Animal vocalizations encode rich biological information—such as age, sex, behavioural context and emotional state—making bioacoustic analysis a promising non-invasive method for assessing welfare and population demography. However, traditional bioacoustic approaches, which rely on manually defined acoustic features, are time-consuming, require specialized expertise and may introduce subjective bias. These constraints reduce the feasibility of analysing increasingly large datasets generated by passive acoustic monitoring (PAM). Transfer learning with Convolutional Neural Networks (CNNs) offers a scalable alternative by enabling automatic acoustic feature extraction without predefined criteria. Here, we applied four pre-trained CNNs—two general purpose models (VGGish and YAMNet) and two avian bioacoustic models (Perch and BirdNET)—to African forest elephant (*Loxodonta cyclotis*) recordings. We used a dimensionality reduction algorithm (UMAP) to represent the extracted acoustic features in two dimensions and evaluated these representations across three key tasks: (1) call-type classification (rumble, roar and trumpet), (2) rumble sub-type identification and (3) behavioural and demographic analysis. A Random Forest classifier trained on these features achieved near-perfect accuracy for rumbles, with Perch attaining the highest average accuracy (0.85) across all call types. Clustering the reduced features identified biologically meaningful rumble sub-types—such as adult female calls linked to logistics—and provided clearer groupings than manual classification. Statistical analyses showed that factors including age and behavioural context significantly influenced call variation ($P < 0.001$), with additional comparisons revealing clear differences among contexts (e.g. nursing, competition, separation), sexes and multiple age classes. Perch and BirdNET consistently outperformed general purpose models when dealing with complex or ambiguous calls. These findings demonstrate that transfer learning enables scalable, reproducible bioacoustic workflows capable of detecting biologically meaningful acoustic variation. Integrating this approach into PAM pipelines can enhance the non-invasive assessment of population dynamics, behaviour and welfare in acoustically active species.

Introduction

The acoustic structure of animal vocalizations encodes a wide array of biological information about a caller's identity, sex, age, body size, emotional state and behavioural context (Briefer et al., 2022; Liao et al., 2018; McCordic et al., 2016; Schwartz et al., 2022). The analysis of vocalizations therefore offers a highly promising avenue for the non-invasive assessment of animal welfare and monitoring of population demographics (Boissy & Lee, 2014; Gibb et al., 2019; Mcloughlin et al., 2019). Passive acoustic monitoring (PAM) is a non-invasive bioacoustic approach that uses continuous, long-term audio recordings collected in natural habitats to monitor wildlife presence, behaviour and ecosystem dynamics (Wrege et al., 2017). PAM uses automated detection algorithms to identify and count vocalizations, providing estimates of population sizes and tracking changes in species activity across time and space. While extracting biological information from the acoustic structure of recorded signals can increase the breadth of insights gained from PAM (Wood et al., 2021), scalability issues remain a significant challenge, particularly as data volumes increase (Gibb et al., 2019; Napier et al., 2024), thereby limiting the scope of application of PAM.

Traditional bioacoustic methods typically involve manually defining and measuring acoustic features (Erbe & Thomas, 2022). These features, such as frequency, duration and amplitude, have been successfully used to establish species-specific call repertoires (e.g. *Orcinus orca*—Selbmann et al., 2023; *Gorilla gorilla*—Salmi et al., 2013) and to correlate call types with biological variables like age, sex, behaviour and emotion (Marler, 1976). However, the manual nature of feature selection is time-consuming, prone to subjective interpretation and limits the ability to process larger datasets (Brown et al., 2018; Janik, 1999; Nguyen Hong Duc et al., 2021). This reliance on manual processes also necessitates specialized bioacoustic expertise, slowing the speed and accessibility of data analysis (Hasan, 2022). Furthermore, disagreements over the classification and definition of features can undermine reproducibility and generalizability (Nguyen Hong Duc et al., 2021; Rekdahl et al., 2013). As a result, these traditional methods are not well suited to the growing demands of large-scale bioacoustic research, particularly in PAM workflows.

Machine learning methods are increasingly used to analyse animal vocalizations. For example, supervised deep learning, in which neural networks are trained to map input data to known output labels, has been used to automatically identify *Orcinus orca* calls (Bergler et al., 2019). Convolutional Neural Networks (CNNs), a specialized class of supervised deep learning, excel in these tasks by automatically learning and extracting features from visual data like audio spectrograms or image pixels (Goffinet et al., 2021). CNNs trained on extensive audio datasets learn to identify and extract audio features from this larger dataset, and these learned acoustic traits can be applied to different audio datasets in a process known as transfer learning (Dufourq et al., 2022; Sethi et al., 2020). Transfer learning using pre-trained CNNs offers several advantages over training custom CNNs from scratch, including reduced computational costs, faster implementation and improved performance on small datasets (Weiss et al., 2016; Yosinski et al., 2014). Pre-trained CNNs also provide deterministic outputs when given the same input, enabling reproducible results (Decuyper et al., 2018). The unsupervised extraction of acoustic features using transfer learning avoids the subjective selection of features required in manual approaches, which can limit analyses to pre-defined traits. Instead, transfer learning uses rich representations learned from diverse datasets to automatically capture complex and subtle acoustic patterns, enhancing the scalability, objectivity and reproducibility of bioacoustic analyses (Sethi et al., 2020; Stowell, 2022). Acoustic features are often numerous and high-dimensional, necessitating dimensionality reduction for efficient analysis and visualization (Jia et al., 2022). Unsupervised dimensionality reduction techniques can simplify data while preserving complex, non-linear relationships, retaining both local and global structures (Ayesha et al., 2020; McInnes et al., 2020). Together, these approaches minimize observer bias in manual methods and improve the scalability and automation of bioacoustic workflows.

Elephant vocalizations are an excellent example of a complex vocal communication system, with all three species displaying a range of calls including the characteristic and most commonly used vocalization—the rumble—as well as a range of broadband vocalizations, such as roars, and trumpets (de Silva, 2010; Hedwig et al., 2019; Poole, 2011). Rumbles are tonal, harmonically rich and often feature fundamental frequencies in the infrasound range (Poole, 2011). In contrast, roars and nasally produced trumpets are higher-frequency broadband calls with noisy but sometimes tonal components (Poole, 2011). Elephants also produce combinatorial calls, where rumbles are combined with broadband calls (mainly roars, but also barks and cries) (Hedwig & Kohlberg, 2024; Pardo et al., 2019). Research on elephant vocalizations has primarily focused on rumbles using traditional bioacoustic methods, producing diverse call-type and sub-type classifications (Leong et al., 2003; Hedwig et al., 2019; Hedwig et al., 2021; Nair et al., 2009; Stoeger et al., 2012, 2021; Wood et al., 2005). The highly graded structural variation in rumbles complicates classification and has been associated with a wide range of biological contexts, including age, sex, behaviour and emotional state (de Silva, 2010; Hedwig et al., 2021; Leong et al., 2003; Nair et al., 2009; Poole, 2011; Stoeger, 2021; Stoeger & de Silva, 2014). Key questions remain about the structure and context of elephant vocalizations, such as the existence of meaningful rumble sub-types. A study using manual feature selection, principal component analysis, and model-based clustering identified eight rumble sub-types, though these sub-types exhibited substantial overlap, underscoring the difficulty of distinguishing clear categories within graded vocalizations (Hedwig et al., 2021). Addressing these gaps would enhance interpretative understanding of elephant calls, enabling more effective non-invasive monitoring of populations and welfare as well as improved demographic assessments for conservation planning based on PAM (Brickson et al., 2023; Wrege et al., 2017).

In this study, we use a transfer learning workflow with four different CNNs to extract acoustic features from African forest elephant (*Loxodonta cyclotis*) vocalizations and evaluate how well these features capture call structure and context. Our methodology involves two stages. First, we separately use the CNNs to automatically extract acoustic features from the vocalizations, generating rich acoustic embeddings for the same dataset. We include both general-purpose (VGGish, YAMNet) and specialized bioacoustic (Perch, BirdNET) CNNs to test whether models trained on broad versus domain-specific datasets yield comparable or distinct insights. Second, we apply dimensionality reduction to these embeddings using Uniform Manifold Approximation and Projection (UMAP), enabling the projection of the embeddings without prior labelling. We then evaluate the effectiveness of the projected acoustic embeddings through three analyses: (1) classification of different call types (roars, rumbles and trumpets), (2) identification of structural variations within rumbles to distinguish sub-types and (3) interpretation of differences in rumbles related to behavioural and demographic factors such as age, sex, behaviour or distress. By comparing the outputs from the different CNN models, we assess the consistency of the embeddings and compare overall model performance. This process enables us to determine whether transfer learning-based feature extraction offers an automatable, standardizable and scalable approach to extracting biologically relevant information from bioacoustic data.

Methods

All code was written in Python (version 3.12.4).

Data collation

We used a labelled dataset of 787 forest elephant (*L. cyclotis*) vocalizations from Hedwig et al. (2021) recorded at Dzanga-Bai in Dzanga-Ndoki National Park (2.963° N, 16.365° E), Central African Republic, between September 2018 and April 2019. Recordings were made with an Earthworks omnidirectional microphone at 48,000 Hz and downsampled to 4,000 Hz for manageability. This preserved relevant frequencies, as most elephant vocalization energy falls below 4 kHz (Poole, 2011). File durations ranged from 0.5 to 39 seconds and typically contained a single call type (rumble, roar, trumpet) or combination calls (e.g. rumble–roar–rumble). Contextual metadata, including the caller's age, sex, behavioural context and distress status, were recorded during visual observations (Hedwig et al., 2021).

The start and end times and frequency ranges of each vocalization were manually annotated using Whombat (Martínez Balvanera et al., 2025). Combination calls were separated into their individual components for call-type classification based on the typically abrupt frequency shifts between roar and rumble components. For example, in rumble-roar-rumble sequences, bounding boxes were drawn manually to define the start and end times and frequency ranges of the rumble and roar segments, which were then relabelled according to their respective call types. This segmentation increased the sample size for each call type. The final dataset contained 1,254 vocalizations: rumbles (n = 779), roars (n = 424) and trumpets (n = 51).

The sound files were labelled by Hedwig et al. (2021) and contained the following contextual labels: sex—female (n = 207), male (n = 107); age—adult (n = 96), infant (n = 14), juvenile (n = 89), sub-adult (n = 129); behaviour context—affiliation (n = 59), competition (n = 79), logistics (n = 15), nursing (n = 24), separation (n = 146) and distress context—distress (n = 45), no distress (n = 96), other (n = 218). The other values for distress comprised both unknown labels—where the distress status could not be determined—and not applicable labels, which referred to contexts where distress could not occur due to the nature of the behaviour. These were retained to maintain sample size and to provide a contrast to the categories of interest: distress and no distress.

Audio pre-processing

Audio files were converted into mel-spectrograms (Fig. 1A) using the librosa package (version 0.10.2) (McFee et al., 2015). A Butterworth bandpass filter was applied to each vocalization using SciPy (version 1.14.0) to remove frequencies outside annotated ranges (Fig. 1B). We extracted only the relevant time periods of each vocalization (Fig. 1C) and added periods of silence (zero-padding) to each vocalization to ensure it met the CNN input window size (Fig. 1D). The peak amplitude was normalized to a 0–1 scale to avoid amplitude-based biases, such as caller distance to microphone (Fig. 1E). A final bandpass filter removed artefacts introduced during pre-processing (Fig. 1F).

CNN feature extraction

We extracted acoustic features using four pre-trained CNNs: VGGish, YAMNet (Gemmeke et al., 2017; Hershey

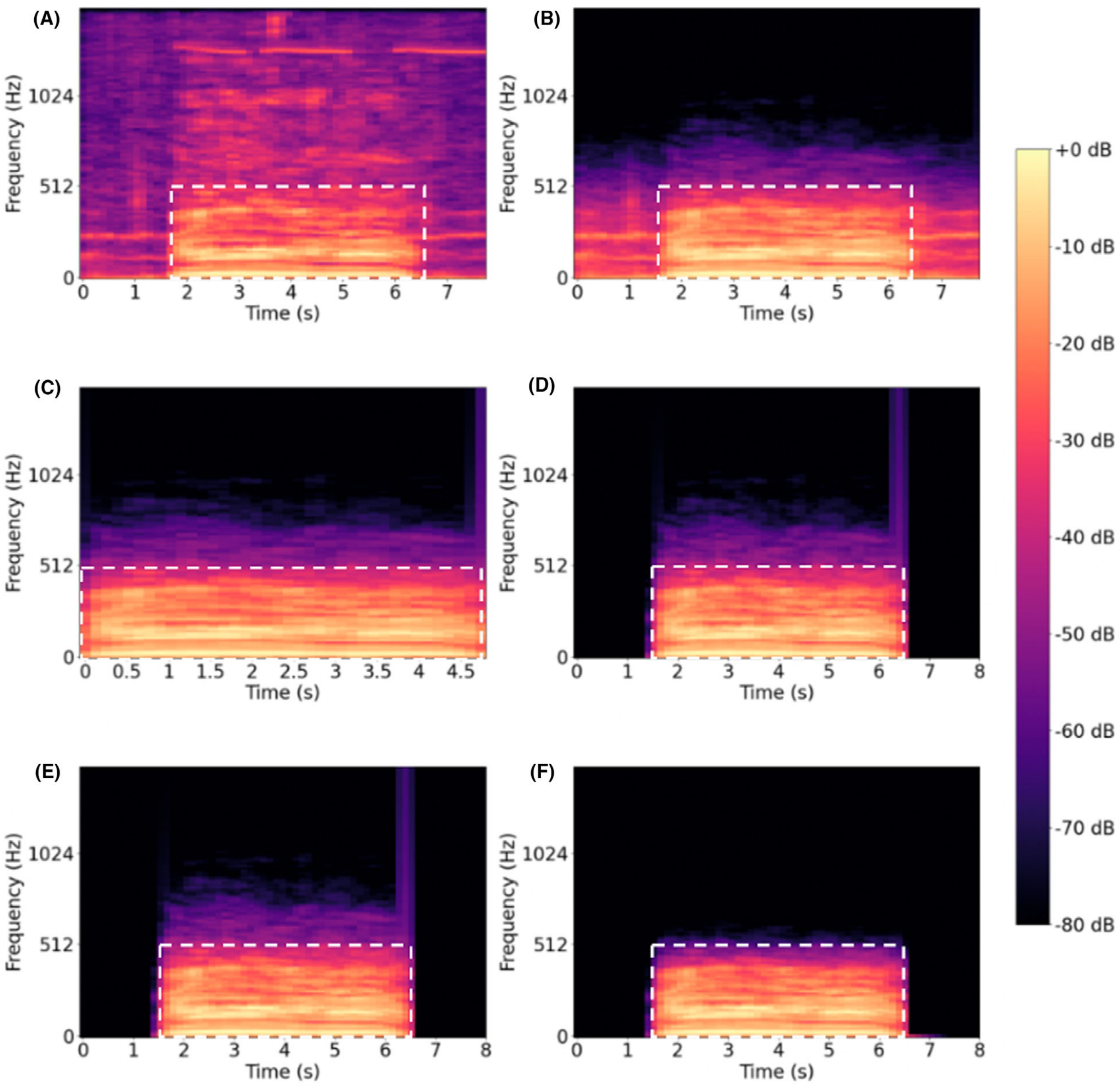


Figure 1. Example mel-spectrogram representations of a single audio file after each audio pre-processing step is applied for the VGGish model. Shading represents amplitude (dB). White dashed bounding box indicates the vocalization. (A) Convert audio into mel-spectrogram, (B) apply bandpass filter, (C) extract relevant time period, (D) zero pad to input window size and centre, (E) normalize amplitude, (F) re-apply bandpass filter.

et al., 2017), Perch (Google Research, 2023) and BirdNET (Kahl et al., 2021). VGGish and YAMNet are general-purpose audio models, both trained on AudioSet—a large-scale dataset comprising over 2 million 10-second audio clips from diverse sources such as YouTube, containing speech, music, environmental sounds and some animal vocalizations (Gemmeke et al., 2017). VGGish, based on the VGG architecture with stacked convolutional layers, was designed to generate broad audio representations for a range of tasks such as music and speech detection, environmental sound recognition, and anomaly detection in audio streams (Hershey et al., 2017). YAMNet ('Yet another Audio Mobilenet Network') is a lightweight model with a MobileNet-inspired architecture optimized for environmental sound recognition, such as detecting birdsong, thunderstorms and ambient noises, making it suitable for real-time applications (Gemmeke et al., 2017).

Perch and BirdNET, by contrast, were trained on extensive avian bioacoustic datasets, including the Xeno-Canto and Macaulay Library collections (Ghani et al., 2023; Kahl et al., 2021) making them specialized for bird vocalization detection. Perch processes longer audio windows to capture detailed acoustic patterns, reflecting its architecture's emphasis on temporal context (Ghani et al., 2023). BirdNET uses a ResNet-inspired architecture designed for large-scale species-level bird identification with high accuracy (Kahl et al., 2021). Despite their avian focus, Perch and BirdNET have demonstrated strong performance in cross-taxa bioacoustic classification tasks (Ghani et al., 2023). This contrast between general-purpose CNNs, designed to extract broad audio features across multiple domains, and bioacoustic-specific CNNs, optimized for detecting detailed acoustic patterns in complex wildlife soundscapes, enabled us to evaluate whether a general or domain-specific model more effectively captured biologically relevant features in elephant vocalizations.

VGGish and YAMNet process 0.96-second spectrogram inputs sampled at 16,000 Hz. To match this input format, we time-expanded the original 4,000 Hz recordings such that 3.84 seconds corresponded to 0.96 seconds at the expanded rate. VGGish outputs 128-dimensional (128D) feature embeddings, while YAMNet outputs 1,024-dimensional (1024D) embeddings. Perch and BirdNET require 5-second and 3-second audio inputs sampled at 32,000 Hz and 48,000 Hz, respectively. Maintaining the original 4,000 Hz sampling rate, we time-expanded the recordings so that the 5-second and 3-second inputs corresponded to 40 seconds and 36 seconds at the expanded rate for Perch and BirdNET, respectively. Perch produces 1,280-dimensional (1280D) embeddings, whereas BirdNET produces 1,024-dimensional (1024D) embeddings. For vocalizations exceeding the input windows, we split the audio into multiple windows, generated individual embeddings for each segment, and averaged these into a single embedding per call. As call duration has been shown to encode biological information (Hedwig et al., 2019; Soltis et al., 2011), we manually reintegrated this information as an additional feature. The final embedding dimensions were 129D (VGGish), 1025D (YAMNet and BirdNET) and 1281D (Perch).

Dimensionality reduction

We applied Uniform Manifold Approximation and Projection (UMAP) for non-linear dimensionality reduction, projecting these high-dimensional embeddings into two dimensions (2D) (McInnes et al., 2020). UMAP was chosen for its superior ability to preserve both local and global structures compared to t-SNE (Becht et al., 2019) and for its faster processing times (Pal & Sharma, 2020). Embeddings were normalized (mean: 0, standard deviation: 1) and projected into 2D using the umap-learn package (version 0.2.0), with cosine distance selected due to its better suitability for high-dimensional data (Ertöz et al., 2003).

For behavioural and demographic analyses, we further reduced the 2D UMAP projections to a single dimension using principal component analysis (PCA) via scikit-learn (version 1.5.2). We selected PCA for this step because it preserves the variance captured by UMAP's non-linear projections while reducing dimensionality through an optimal linear transformation (Jolliffe & Cadima, 2016). Importantly, the use of PCA addresses the challenge of small sample sizes, as fitting more complex models to 2D embeddings would risk overfitting. The resulting single principal component serves as an acoustic index, representing the primary variation in acoustic features for each rumble vocalization in a simplified yet informative way.

Modelling

Call-type classification

We assessed how well the 2D UMAP projections retained call-type information by evaluating clustering with silhouette scores and training a random forest (RF) classifier to predict the call types. Silhouette scores quantify cluster separation by comparing the mean distance between points within the same cluster (a) to points in the nearest cluster (b) Rousseeuw (1987). Overall and per-call-type scores were calculated, with ≥0.5 indicating good clustering and ≥0.7 indicating strong clustering (Dalmaijer et al., 2022). To support interpretability, we visualized the 2D UMAP projections and generated spectrograms from a subsample of the projections for the best-performing model as an example of the acoustic differences between call types, following the approach described by Sainburg et al. (2020).

The RF model was optimized using GridSearchCV (scikit-learn, version 1.5.2) to tune class weights, tree depth, minimum leaf and split samples and estimators, balancing complexity and overfitting risks (Probst et al., 2019). Minority classes were oversampled using RandomOverSampler (imblearn, version 0.12.3). Nested cross-validation (CV) minimized tuning bias by using inner CV loops for parameter optimization and outer CV loops for error estimation (Varma & Simon, 2006). Performance was evaluated using macro average accuracy and F1 scores, which reflect balanced performance across classes and capture the trade-off between precision and recall (Powers, 2020).

Call sub-type identification

To identify rumble subtypes, we used Affinity Propagation Clustering (APC) on the 2D UMAP projections.

Unlike traditional clustering algorithms that require the user to specify the number of clusters beforehand, APC identifies exemplars within the data based on similarity measures and iteratively updates cluster assignments until convergence (Frey & Dueck, 2007). This allows APC to automatically determine the number of clusters and capture complex data structures with varying cluster sizes and densities. We used the AffinityPropagation package in scikit-learn (version 1.5.1) with preference parameters set to −90 (to avoid exemplar bias) and damping to 0.95 (to stabilize convergence) (Frey & Dueck, 2007). Silhouette scores were used to assess clustering performance, consistent with the call-type analysis. To characterize each cluster, we calculated the proportions of levels within each category (e.g. age categories: adult, juvenile and infant) for each cluster and model. This allowed us to identify which levels (e.g. adult age class) were dominant within a cluster, enabling visual inspection of the key demographic and behavioural patterns.

Behavioural and demographic analysis

We fitted generalized linear models (GLMs) with normal error structures and identity link functions (statsmodels, version 0.14.0) to test for differences in acoustic embeddings across age, sex, behavioural context and distress status. The dependent variable was the 1D PCA projection of the acoustic embeddings for each model, representing the primary dimension of acoustic variation. Predictor variables included age (adult, sub-adult, juvenile, infant), sex (male, female), behavioural context (affiliation, competition, logistics, nursing and separation) and distress status (yes, no and other). GLM fit was assessed using standard performance metrics: the coefficient of determination (R^2) for explanatory power, mean absolute error (MAE) for average deviation from true values and root mean squared error (RMSE) to emphasize larger prediction errors. The significance of each predictor category was assessed using likelihood ratio tests (LRTs), comparing the full model to reduced models excluding each predictor category one at a time. Tukey post hoc pairwise comparisons (SciPy, version 1.14.0) were conducted to determine which specific levels of the predictor variables differed significantly ($P < 0.05$), controlling for multiple comparisons.

Results

Call-type clustering and classification

Silhouette scores across all models exceeded the 0.5 threshold, indicating effective call-type clustering (Table 1). BirdNET achieved the highest silhouette score (0.59), demonstrating the strongest overall clustering performance (Fig. 2). It performed particularly well for rumbles, achieving a silhouette score of 0.89. In contrast, trumpets exhibited weaker clustering (0.35), and roars were poorly clustered, with BirdNET achieving a silhouette score of only 0.08. Other models showed slightly lower overall silhouette scores: Perch and VGGish scored 0.57, while YAMNet scored 0.55 (Table 1; Fig. S1). Rumbles clustered strongly across all models, with silhouette scores ranging from 0.82 (VGGish, YAMNet) to 0.89 (BirdNET). Trumpets displayed greater variability, with Perch achieving a silhouette score of 0.37 and VGGish scoring just 0.01. Roars consistently showed weak clustering, with scores ranging from 0.08 (BirdNET, YAMNet) to 0.18 (VGGish).

Table 1. Call-type clustering and classification results comparing the four CNNs.

	VGGish	Perch	YAMNet	BirdNET
Call-type clustering				
Rumble—Silhouette score	0.82	0.84	0.82	0.89
Roar—Silhouette score	0.18	0.10	0.08	0.08
Trumpet—Silhouette score	0.01	0.37	0.25	0.35
Overall—Silhouette score	0.57	0.57	0.55	0.59
Call-type classification				
Rumble—F1 score	1	1	1	1
Roar—F1 score	0.95	0.96	0.94	0.95
Trumpet—F1 score	0.57	0.59	0.50	0.56
Overall accuracy	0.96	0.97	0.96	0.97
Macro average accuracy	0.84	0.85	0.81	0.84

Random forest classifiers demonstrated strong macro average accuracy (MAA) across the CNN models. Perch achieved the highest MAA (0.85), followed by BirdNET and VGGish (0.84 each), and YAMNet (0.81). Rumbles were classified with near-perfect accuracy across all models (F1 score: 1.0). Roars achieved high F1 scores ranging from 0.94 (YAMNet) to 0.96 (Perch). Trumpets remained the most challenging call type, with F1 scores between 0.50 (YAMNet) and 0.59 (Perch) (Table 1).

Call sub-type identification

The affinity propagation unsupervised clustering identified between 5 and 6 rumble sub-types across the four CNN models, with silhouette scores indicating moderate clustering performance (range: 0.46–0.51) (see Fig. S2). YAMNet achieved the highest silhouette score (0.51) and identified 5 clusters, with VGGish, and Perch also identifying 5 clusters with lower average silhouette scores of 0.50 and 0.46, respectively. BirdNET identified 6 clusters with an average silhouette score of 0.50. There was insufficient data to perform statistical tests to determine the

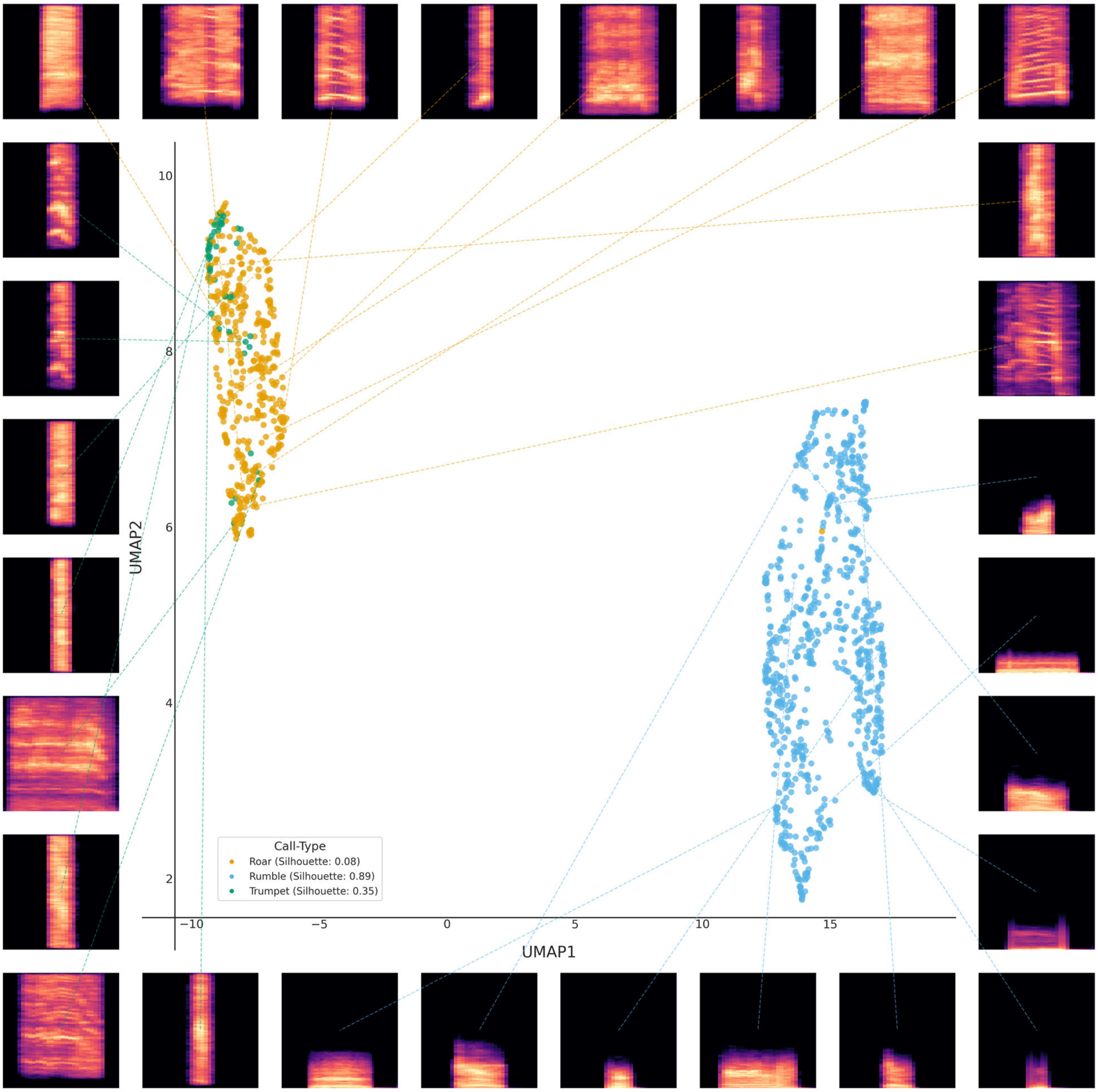


Figure 2. UMAP projections of acoustic features of elephant call types into 2D space using BirdNET CNN. 2D UMAP projections of the acoustic embeddings for forest elephant roars, rumbles and trumpets for the best-performing CNN for call-type clustering, BirdNET. Each point represents a single vocalization. Colour indicates call type. Silhouette scores are shown per call type in the legend. Sample points show underlying spectrogram images of vocalizations.

composition of the clusters by biological category; however, visual inspection of the distribution of category proportions within the clusters suggests biologically relevant differences (Fig. 3).

For example, YAMNet produced a cluster (cluster 3) containing a high proportion of adult female calls associated with the logistics context (53% of all adult vocalizations, 28% of female vocalizations, and 30% of logistics-related calls) and another cluster (cluster 4) predominantly comprising juvenile male calls associated with nursing (43% juvenile, 37% male and 52% nursing-related calls).

Corresponding clusters were observed across the other models, indicating consistency across CNN architectures (Fig. S3). For the putative adult female logistics cluster identified by YAMNet, VGGish produced a corresponding

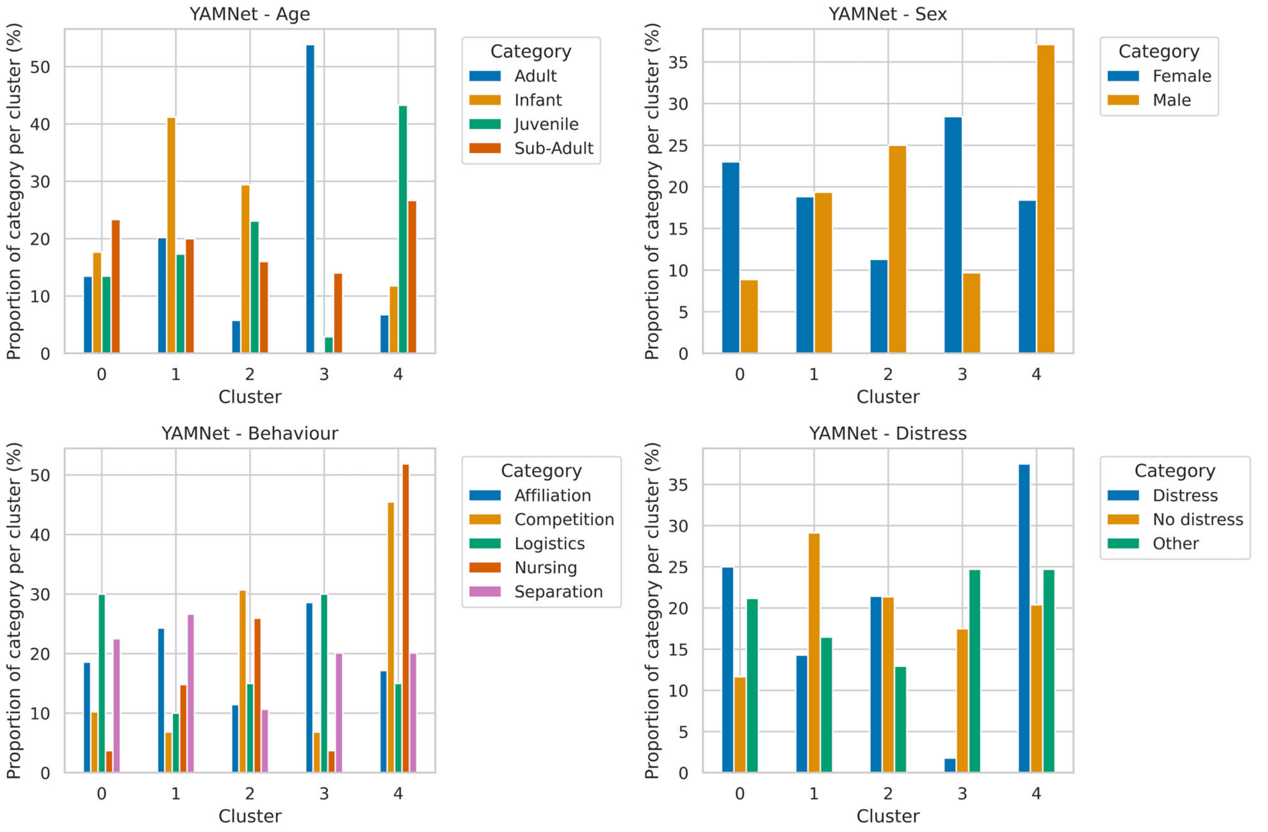


Figure 3. Bar chart of rumbles with biological labels and cluster classification for YAMNet CNN. *X*-axis shows cluster number, and y-axis shows percentage proportion of a category label (e.g. 'Infant') distributed across all clusters. Quadrants represent different categories: age, sex, behaviour, distress).

cluster (cluster 3) with 62% adult, 33% female and 35% logistics-related calls, while Perch identified a similar cluster (cluster 4) with 52% adult, 28% female and 40% logistics calls. Similarly, for the tentative juvenile male nursing cluster identified by YAMNet (cluster 4), VGGish identified a corresponding cluster (cluster 2) with 38% juvenile, 36% male and 63% nursing-related calls, while Perch identified cluster 3 with 47% juvenile, 47% male and 70% nursing calls.

Behavioural and demographic analysis

Results from the GLMs showed that Perch had the best fit to the data, with the lowest mean absolute error (MAE: 1.34) and the highest R^2 value (0.47), indicating strong explanatory power (Table 2). BirdNET also performed well (MAE: 1.38, R^2: 0.42), whereas YAMNet (MAE: 1.41, R^2: 0.38) and VGGish (MAE: 1.58, R^2: 0.34) showed more moderate performances.

Behaviour and age were both significant predictors across all models ($P < 0.001$), indicating that the acoustic features of rumbles varied significantly across both contexts (Table 2 and see Supporting Information, Table S1 for full GLM model coefficients for each CNN). Distress was a significant predictor for VGGish ($P < 0.001$), Perch and YAMNet ($P < 0.01$), but not for BirdNET ($P = 0.1394$). Sex was only significant for BirdNET ($P < 0.05$), though post hoc pairwise comparisons

Table 2. Summary table of GLMs per CNN comparing model performance and predictor category impact on elephant rumble vocalizations.

Model	MAE	RMSE	R^2	Behaviour	Distress	Age	Sex
VGGish	1.58	1.92	0.34	0.0000***	0.0002***	0.0000***	0.5828
Perch	1.34	1.75	0.47	0.0000***	0.0018**	0.0000***	0.1972
YAMNet	1.41	1.76	0.38	0.0000***	0.0049**	0.0000***	0.0601
BirdNET	1.38	1.71	0.42	0.0000***	0.1394	0.0000***	0.0309*

P-value significance codes: *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

showed distinctions between male and female calls for all models.

The Tukey post hoc pairwise comparisons provided further detail on the significant differences between behavioural categories (Fig. 4). Competition calls were significantly different from all other behavioural contexts across all models ($P < 0.01$), except for nursing, and nursing calls were significantly different from all other contexts ($P < 0.001$) except for competition. VGGish was the only model to detect differences between nursing and competition ($P < 0.05$). Perch and BirdNET distinguished calls made in a logistics context from separation ($P < 0.05$). For age, all models identified significant differences between adult vocalizations and all other age classes ($P < 0.001$). Perch, YAMNet and BirdNET further distinguished sub-adults from juveniles ($P < 0.001$) and sub-adults from infants (Perch and BirdNET $P < 0.01$; YAMNet $P < 0.05$). VGGish and YAMNet were the only models to detect differences between distress and non-distress contexts ($P < 0.05$). For sex, significant differences between male and female calls were detected across all models ($P < 0.001$) (Fig. 4).

Discussion

The manual bioacoustic analysis of animal vocalizations poses several practical challenges, including time-consuming processing, subjective interpretation and limited scalability (Brown et al., 2018; Janik, 1999; Nguyen Hong Duc et al., 2021). These challenges, particularly in feature extraction, highlight the importance of developing scalable, automated workflows in bioacoustic research, especially for large-scale passive acoustic monitoring (PAM). By applying transfer learning methods to African forest elephant vocalizations, we established a robust, modular workflow encompassing audio preprocessing, CNN feature extraction, dimensionality reduction and classification of acoustic outputs. Our workflow provides valuable insights into age and sex differences, behavioural contexts and distress levels, improving our ability to interpret social dynamics and potential stress responses. Its scalability and reproducibility make it well suited for large-scale PAM, enabling more efficient tracking of demographic trends and welfare indicators in forest elephants and other species.

Call-type classification

All CNNs demonstrated strong agreement in distinguishing call types, particularly for rumbles, which clustered consistently across all models. This shows that CNN-based embeddings effectively captured the defining acoustic features of rumbles, even when derived from different training backgrounds. However, performance varied for more ambiguous call types. Trumpets were more challenging to classify, with bioacoustic CNNs (BirdNET and Perch) outperforming general-purpose models in identifying key acoustic patterns. The overall classification performance for forest elephant calls compares favourably with other bioacoustic studies. For example, our RF classifier achieved a best overall accuracy of 0.97 (BirdNET and Perch), which exceeds the 0.83 accuracy reported for a comparable classifier used to categorise Asian elephant call types (Lokhandwala et al., 2023). These findings highlight the potential of CNN-based embeddings to robustly classify complex vocal repertoires and support the efficient identification of forest elephant call types from passive acoustic monitoring data.

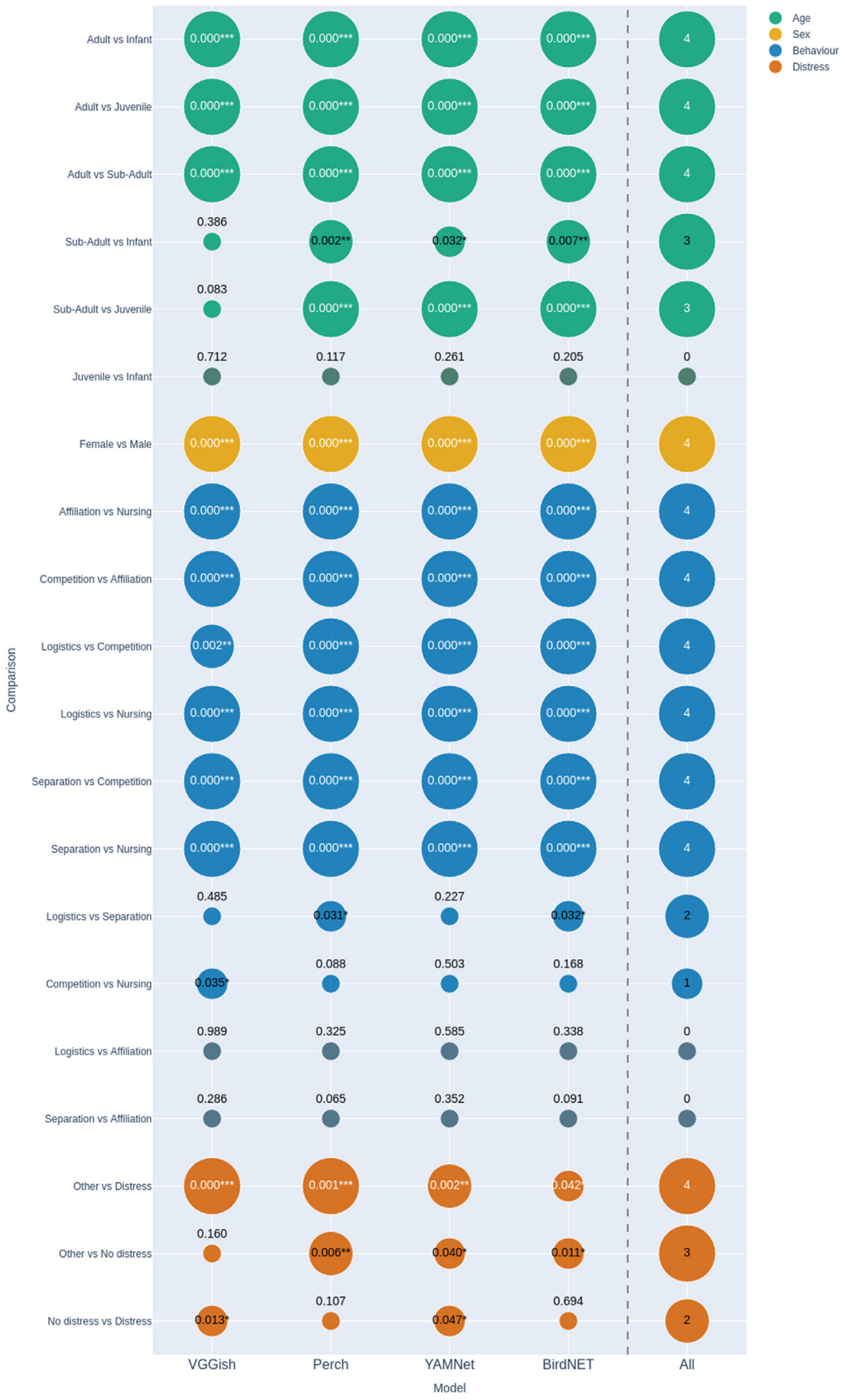
Call sub-type identification

In the unsupervised clustering of rumbles, the CNN models identified 5–6 sub-types, with silhouette scores indicating moderate clustering performance. YAMNet achieved the highest silhouette score (0.51) and three models (VGGish, Perch and YAMNet) consistently identified clusters with biological relevance, such as those associated with adult female logistics calls and juvenile male nursing calls. These findings suggest that the models captured meaningful variations in vocalizations linked to demographic and behavioural factors. However, it is important to note that these are tentative initial descriptions and are based on differentiating clusters through proportions of context. Furthermore, some categories such as separation are present in all clusters and sample sizes for several levels are small.

The performance of the transfer learning approach in identifying rumble sub-types also outperformed traditional clustering methods. Hedwig et al. (2019) identified eight rumble sub-types using a combination of manual feature selection, principal component analysis and model-based clustering, with average silhouette coefficients below 0.34. In contrast, our overall silhouette scores (range 0.46–0.51) indicate a more robust separation of sub-types, demonstrating the potential of automated feature extraction methods to provide meaningful acoustic clustering without requiring domain-specific manual feature selection.

Behavioural and demographic analysis

The GLM analysis confirmed that behavioural context and age were significant predictors of acoustic variation across all models. This is consistent with previous findings that competition and nursing contexts produce



Age
Sex
Behaviour
Distress
Comparison
Adult vs Infant 0.000*** 0.000*** 0.000*** 0.000*** 4
Adult vs Juvenile 0.000*** 0.000*** 0.000*** 0.000*** 4
Adult vs Sub-Adult 0.000*** 0.000*** 0.000*** 0.000*** 4
Sub-Adult vs Infant 0.386 0.002** 0.032* 0.007** 3
Sub-Adult vs Juvenile 0.083 0.000*** 0.000*** 0.000*** 3
Juvenile vs Infant 0.712 0.117 0.261 0.205 0
Female vs Male 0.000*** 0.000*** 0.000*** 0.000*** 4
Affiliation vs Nursing 0.000*** 0.000*** 0.000*** 0.000*** 4
Competition vs Affiliation 0.000*** 0.000*** 0.000*** 0.000*** 4
Logistics vs Competition 0.002** 0.000*** 0.000*** 0.000*** 4
Logistics vs Nursing 0.000*** 0.000*** 0.000*** 0.000*** 4
Separation vs Competition 0.000*** 0.000*** 0.000*** 0.000*** 4
Separation vs Nursing 0.000*** 0.000*** 0.000*** 0.000*** 4
Logistics vs Separation 0.485 0.031* 0.227 0.032* 2
Competition vs Nursing 0.035* 0.088 0.503 0.168 1
Logistics vs Affiliation 0.989 0.325 0.585 0.338 0
Separation vs Affiliation 0.286 0.065 0.352 0.091 0
Other vs Distress 0.000*** 0.001*** 0.002** 0.042* 4
Other vs No distress 0.160 0.006** 0.040* 0.011* 3
No distress vs Distress 0.013* 0.107 0.047* 0.694 2
VGGish
Perch
YAMNet
BirdNET
All
Model

Figure 4. Tukey Honestly significant difference (HSD) pairwise tests between all category levels. Each row represents a pair within a category (e.g. Adult vs. Juvenile) and each column is a model, with the fifth column summing how many models identified a significant ($P < 0.05$) Tukey HSD between the paired calls. *P*-value and asterisk labelled for each pair and model, with bubble size indicating significance (larger bubbles = lower *P*-value). The colours represent the four categories. *P*-value significance codes: *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

distinctive acoustic signatures (Hedwig et al., 2021). Distress-related differences, however, were detected inconsistently across models. Only VGGish and YAMNet reliably distinguished distress from non-distress calls, with BirdNET appearing less suited to detecting distress-related differences. These results suggest that the models vary in their sensitivity to subtle acoustic variations associated with emotional states. For sex, BirdNET was the only model to identify this category as a significant predictor of acoustic variation. Post hoc comparisons revealed significant differences between male and female vocalizations across all CNNs.

The ability of the models to extract age and behavioural information aligns with findings from previous studies using manual acoustic analysis. For example, Stoeger et al. (2014) demonstrated that frequency-based features could distinguish between adult and juvenile vocalizations in African elephants, whereas Hedwig et al. (2021) showed that rumbles produced in competition and nursing contexts had shorter durations than those in other behavioural contexts. Our findings show that CNN-based embeddings can autonomously differentiate many of the same categories identified in manual analyses, without the need for feature pre-selection. This highlights the potential for CNNs to offer more flexible and scalable alternatives to manual approaches.

Limitations and future directions

The developed workflow involved manually annotating sound events in audio files to match observed behaviour with the acoustic structure of recorded vocalizations—an inherently labour-intensive and time-consuming process. In PAM studies, this step is increasingly streamlined using semi-automated approaches that apply detection algorithms followed by manual verification (Bjorck et al., 2019; Mcloughlin et al., 2019). While these semi-automated approaches still require human input, they do not demand expert-level knowledge of vocal production or acoustic analysis. Instead, verification can be performed by trained non-experts, making the process more efficient, with the more complex and specialized task of acoustic feature extraction being fully automated within our workflow. However, the need for manual or semi-automated annotation remains a key limitation. Developing fully automated solutions for this task is currently constrained by the scarcity of high-quality labelled datasets for training and validation. For many species, annotated audio data with precise behavioural labels is limited, creating bottlenecks in optimizing detection and classification models for PAM workflows. Expanding datasets to include richer annotations—such as behavioural context and combinatorial vocal sequences—would improve the accuracy and robustness of automated approaches across diverse environments and populations.

Beyond the challenge of dataset expansion, analysing individual calls in isolation limits the ability to capture the full complexity of communication. Combinatorial sequences can convey nuanced information in social contexts (Hedwig & Kohlberg, 2024) and reveal patterns that single calls cannot. Sequence-based deep learning models—such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and attention-based transformers—are well suited for analysing these temporal structures and could significantly improve interpretations of combinatorial vocal sequences (Vaswani et al., 2017).

Although there was strong consensus among the four models for most tasks, minor yet noteworthy differences emerged that appear linked to each model's training corpus. BirdNET and Perch, both trained on avian datasets, performed best at broad call-type classification and identifying behavioural or demographic cues, likely because their filters capture harmonic or temporal structures shared by many bird and elephant calls (Elemans et al., 2015; Ghani et al., 2023). In contrast, VGGish and YAMNet—exposed to more varied audio sources—showed a heightened ability to detect subtle rumble distress signals, potentially reflecting a stronger emphasis on fine-grained acoustic modulations. These findings suggest that even small domain mismatches may confer specialized advantages or limitations. As a result, researchers should carefully consider the alignment between their target sounds and the model's training data when selecting a CNN for transfer learning.

Broader implications

PAM approaches have been successfully used to estimate population sizes and activity patterns from acoustic data, with particular success for cryptic species such as forest elephants (Swider, 2023; Verahrami, 2023; Wrege et al., 2017). Scalable bioacoustic methods to extract biological information from recorded vocalizations would enable widening the scope of PAM applications, with the potential to improve how we monitor populations and behaviour of elusive but acoustically conspicuous species. These results demonstrate that biologically relevant information can be derived from vocalizations using CNN-based embeddings in an automatable, standardizable and scalable way. In particular, the transfer learning approach employed here automated the extraction of features that distinguish adults from juveniles, males from females, distress from non-distress and competition from other behavioural contexts. This automation substantially reduces the manual workload of acoustic analysis, enabling the more efficient processing of large datasets. Applied at a landscape scale, these techniques could facilitate monitoring demographic changes, indirectly track reproductive rates and evaluate the impact of activities such as logging on stress levels. These findings have important implications for evidence-based conservation, as scalable PAM workflows that capture nuanced acoustic information can inform management decisions, improve welfare monitoring, and support long-term conservation efforts for acoustically active species.

Acknowledgements

This research was supported by a Natural Environment Research Council grant to AP (grant number NE/S007229/1), a National Geographic Society (grant #28718 to DH) and a generous gift from K. Lisa Yang to the Cornell Lab of Ornithology. SMB was supported by a CONACYT grant (Grant/Award Number: 2020-000017-02EXTF-00334). We thank the government of the Central African Republic for permission to work in the Dzanga-Ndoki National Park and the Dzanga-Sangha Protected Area Management and WWF-CAR for logistical support. Special thanks go to Saphira Bekoma Mbala for assistance with data collection. AP would like to thank Daniel Maynard, Ella Browning and Rory Gibb for helpful discussions.

Author Contributions

Conceptualization: AP, SMB, KJ, DH. Data curation: AP. Formal analysis: AP. Methodology: AP, SMB. Resources: KJ, DH. Software: AP. Supervision: SMB, KJ, DH. Validation: SMB. Visualization: AP. Writing—original draft preparation: AP. Writing—review and editing: AP, SMB, KJ, DH.

Data Availability Statement

All code needed to run the workflow is available at: www.github.com/AlastairPickering/elephant_transfer_learning. Example audio files are included. The full dataset is available upon request to dh646@cornell.edu.

References

Ayesha, S., Hanif, M.K. & Talib, R. (2020) Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, **59**, 44–58. Available from: https://doi.org/10.1016/j.inffus.2020.01.005

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G. et al. (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, **37**(1), 38–44. Available from: https://doi.org/10.1038/nbt.4314

Bergler, C., Schröter, H., Cheng, R.X., Barth, V., Weber, M., Nöth, E. et al. (2019) ORCA-SPOT: an automatic killer whale sound detection toolkit using deep learning. *Scientific Reports*, **9**(1). Available from: https://doi.org/10.1038/s41598-019-47335-w

Bjorck, J., Rappazzo, B.H., Chen, D., Bernstein, R., Wrege, P.H. & Gomes, C.P. (2019) Automatic detection and compression for passive acoustic monitoring of the African Forest elephant. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(1), 476–484. Available from: https://doi.org/10.1609/aaai.v33i01.3301476

Boissy, A. & Lee, C. (2014) How assessing relationships between emotions and cognition can improve farm animal welfare. *Revue Scientifique et Technique – Office International des Épizooties*, **33**(1), 103–110.

Brickson, L., Zhang, L., Vollrath, F., Douglas-Hamilton, I. & Titus, A.J. (2023) Elephants and algorithms: a review of the current and future role of AI in elephant monitoring. *Journal of the Royal Society Interface*, **20**(208), 20230367. Available from: https://doi.org/10.1098/rsif.2023.0367

Briefer, E.F., Sypherd, C.C.-R., Linhart, P., Leliveld, L.M.C., Padilla de la Torre, M., Read, E.R. et al. (2022) Classification of pig calls produced from birth to slaughter according to their emotional valence and context of production. *Scientific Reports*, **12**(1), 3409. Available from: https://doi.org/10.1038/s41598-022-07174-8

Brown, A., Garg, S. & Montgomery, J. (2018) Scalable preprocessing of high volume environmental acoustic data for bioacoustic monitoring. *PLoS One*, **13**(8), e0201542. Available from: https://doi.org/10.1371/journal.pone.0201542

Dalmaijer, E.S., Nord, C.L. & Astle, D.E. (2022) Statistical power for cluster analysis. *BMC Bioinformatics*, **23**(1), 205. Available from: https://doi.org/10.1186/s12859-022-04675-1

de Silva, S. (2010) Acoustic communication in the Asian elephant, Elephas maximus maximus. *Behaviour*, **147**(7), 825–852. Available from: https://doi.org/10.1163/000579510X495762

Decuyper, M., Bonte, S. & Van Holen, R. (2018) *Binary glioma grading: radiomics versus pre-trained CNN features*. Granada, Spain: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pp. 498–505. Available from: https://doi.org/10.1007/978-3-030-00931-1_57

Dufourq, E., Batist, C., Foquet, R. & Durbach, I. (2022) Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics*, **70**, 101688. Available from: https://doi.org/10.1016/j.ecoinf.2022.101688

Elemans, C.P.H., Rasmussen, J.H., Herbst, C.T., Düring, D.N., Zollinger, S.A., Brumm, H. et al. (2015) Universal mechanisms of sound production and control in birds and mammals. *Nature Communications*, **6**(1), 8978. Available from: https://doi.org/10.1038/ncomms9978

Erbe, C. & Thomas, J.A. (Eds.). (2022) *Exploring animal behavior through sound: volume 1: methods*. Cham: Springer International Publishing. Available from: https://doi.org/10.1007/978-3-030-97540-1

Ertöz, L., Steinbach, M. & Kumar, V. (2003) Finding clusters of different sizes, shapes, and densities in Noisy, high dimensional data. In: *Proceedings of the 2003 SIAM international conference on data mining (SDM)*. Society for Industrial and Applied Mathematics (Proceedings), San Francisco, CA, pp. 47–58. Available from: https://doi.org/10.1137/1.9781611972733.5

Frey, B.J. & Dueck, D. (2007) Clustering by passing messages between data points. *Science*, **315**(5814), 972–976. Available from: https://doi.org/10.1126/science.1136800

Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C. et al. (2017) Audio set: an ontology and human-labeled dataset for audio events. In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, pp. 776–780. Available from: https://doi.org/10.1109/ICASSP.2017.7952261

Ghani, B., Denton, T., Kahl, S. & Klinck, H. (2023) Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, **13**(1), 22876. Available from: https://doi.org/10.1038/s41598-023-49989-z

Gibb, R., Browning, E., Glover-Kapfer, P. & Jones, K.E. (2019) Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, **10**(2), 169–185. Available from: https://doi.org/10.1111/2041-210X.13101

Goffinet, J., Brudner, S., Mooney, R. & Pearson, J. (2021) Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *eLife*, **10**, e67855. Available from: https://doi.org/10.7554/elife.67855

Google Research. (2023) Google bird vocalization classifier: a global bird embedding and classification model. Available at: https://www.kaggle.com/models/google/bird-vocalization-classifier [Accessed 4 January 2025]

Hasan, N.I. (2022) Bird species classification and acoustic features selection based on distributed neural network with two stage windowing of short-term features. *arXiv*. Available at: http://arxiv.org/abs/2201.00124 [Accessed 3 August 2023]

Hedwig, D. & Kohlberg, A. (2024) Call combination in African forest elephants *Loxodonta cyclotis*. *PLoS One*, **19**(3), e0299656. Available from: https://doi.org/10.1371/journal.pone.0299656

Hedwig, D., Poole, J. & Granli, P. (2021) Does social complexity drive vocal complexity? Insights from the two African elephant species. *Animals*, **11**(11), 3071. Available from: https://doi.org/10.3390/ani11113071

Hedwig, D., Verahrami, A.K. & Wrege, P.H. (2019) Acoustic structure of forest elephant rumbles: a test of the ambiguity reduction hypothesis. *Animal Cognition*, **22**(6), 1115–1128. Available from: https://doi.org/10.1007/s10071-019-01304-y

Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C. et al. (2017) CNN architectures for large-scale audio classification. *arXiv*. Available at: http://arxiv.org/abs/1609.09430 [Accessed 12 May 2023]

Janik, V.M. (1999) Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods. *Animal Behaviour*, **57**(1), 133–143. Available from: https://doi.org/10.1006/anbe.1998.0923

Jia, W., Sun, M., Lian, J. & Hou, S. (2022) Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, **8**(3), 2663–2693. Available from: https://doi.org/10.1007/s40747-021-00637-x

Jolliffe, I.T. & Cadima, J. (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **374**(2065), 20150202. Available from: https://doi.org/10.1098/rsta.2015.0202

Kahl, S., Wood, C.M., Eibl, M. & Klinck, H. (2021) BirdNET: a deep learning solution for avian diversity monitoring. *Ecological Informatics*, **61**, 101236. Available from: https://doi.org/10.1016/j.ecoinf.2021.101236

Leong, K.M., Ortolani, A., Burks, K.D., Mellen, J.D. & Savage, A. (2003) Quantifying acoustic and temporal characteristics of vocalizations for a group of captive African elephants Loxodonta Africana. *The International Journal of Animal Sound and its Recording*, **13**, 213–231. Available from: https://doi.org/10.1080/09524622.2003.9753499

Liao, D., Liao, D.A., Zhang, Y.S., Cai, L.X. & Ghazanfar, A.A. (2018) Internal states and extrinsic factors both determine monkey vocal production. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(15), 3978–3983. Available from: https://doi.org/10.1073/pnas.1722426115

Lokhandwala, S., Sinha, R., Ganji, S. & Pailla, B. (2023) Decoding Asian elephant Vocalisations: unravelling call types, context-specific behaviors, and individual identities. In: Karpov, A., Samudravijaya, K., Deepak, K.T., Hegde, R.M., Agrawal, S.S. & Prasanna, S.R.M. (Eds.) *Speech and computer*. Cham: Springer Nature Switzerland (Lecture Notes in Computer Science), pp. 367–379. Available from: https://doi.org/10.1007/978-3-031-48312-7_30

Marler, P. (1976) Social organisation, communication and graded signals: the chimpanzee and the gorilla. In: Bateson, P.P.G. & Hinde, R.A. (Eds.) *Growing points in ethology*. Press Cambridge: Cambridge Univ, pp. 239–265.

Martínez Balvanera, S., Mac Aodha, O., Weldy, M.J., Pringle, H., Browning, E. & Jones, K.E. (2025) Whombat: an open-source audio annotation tool for machine learning assisted bioacoustics. *Methods in Ecology and Evolution*, **16**(1), 19–28. Available from: https://doi.org/10.1111/2041-210X.14468

McCordic, J., Root-Gutteridge, H., Cusano, D.A., Denes, S.L. & Parks, S.E. (2016) Calls of North Atlantic right whales Eubalaena glacialis contain information on individual identity and age class. *Endangered Species Research*, **30**, 157–169. Available from: https://doi.org/10.3354/esr00735

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E. et al. (2015) Librosa: audio and music signal analysis in python. In: Python in science conference, Austin, Texas, pp. 18–24 https://doi.org/10.25080/Majora-7b98e3ed-003

McInnes, L., Healy, J. & Melville, J. (2020) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv* https://doi.org/10.48550/arXiv.1802.03426

Mcloughlin, M.P., Stewart, R. & McElligott, A.G. (2019) Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface*, **16**(155), 20190225. Available from: https://doi.org/10.1098/rsif.2019.0225

Nair, S., Balakrishnan, R., Seelamantula, C.S. & Sukumar, R. (2009) Vocalizations of wild Asian elephants (Elephas maximus): structural classification and social context. *The Journal of the Acoustical Society of America*, **126**(5), 2768–2778. Available from: https://doi.org/10.1121/1.3224717

Napier, T., Ahn, E., Allen-Ankins, S., Schwarzkopf, L. & Lee, I. (2024) Advancements in preprocessing, detection and classification techniques for ecoacoustic data: a comprehensive review for large-scale passive acoustic monitoring. *Expert Systems with Applications*, **252**, 124220. Available from: https://doi.org/10.1016/j.eswa.2024.124220

Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P.R., Gérard, O., Adam, O. et al. (2021) Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. *Ecological Informatics*, **61**, 101185. Available from: https://doi.org/10.1016/j.ecoinf.2020.101185

Pal, K. & Sharma, M. (2020) Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space. In: *2020 fourth international conference on I-SMAC (IoT in social, Mobile, analytics and cloud) (I-SMAC). 2020 fourth international conference on I-SMAC (IoT in social, Mobile, analytics and cloud) (I-SMAC)*, Palladam, India, pp. 1106–1110. Available from: https://doi.org/10.1109/I-SMAC49090.2020.9243502

Pardo, M.A., Poole, J.H., Stoeger, A.S., Wrege, P.H., O'Connell-Rodwell, C.E., Padmalal, U.K. et al. (2019) Differences in combinatorial calls among the 3 elephant species cannot be explained by phylogeny. *Behavioral Ecology*, **30**(3), 809–820. Available from: https://doi.org/10.1093/beheco/arz018

Poole, J. (2011) The behavioral context of African elephant acoustic communication. In: Moss, C.J., Croze, H. & Lee, P.C. (Eds.) *The Amboseli elephants: a long-term perspective on a long-lived mammal*. Chicago: University of Chicago Press, pp. 125–161.

Powers, D.M.W. (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* https://doi.org/10.48550/arXiv.2010.16061

Probst, P., Boulesteix, A.L. & Bischl, B. (2019) Tunability: importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, **20**(53), 1–32.

Rekdahl, M.L., Dunlop, R.A., Noad, M.J. & Goldizen, A.W. (2013) Temporal stability and change in the social call repertoire of migrating humpback whales. *The Journal of the Acoustical Society of America*, **133**(3), 1785–1795. Available from: https://doi.org/10.1121/1.4789941

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**(1), 53–65. Available from: https://doi.org/10.1016/0377-0427(87)90125-7

Sainburg, T., Thielk, M. & Gentner, T.Q. (2020) Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology*, **16**(10), e1008228. Available from: https://doi.org/10.1371/journal.pcbi.1008228

Salmi, R., Hammerschmidt, K. & Doran-Sheehy, D.M. (2013) Western gorilla vocal repertoire and contextual use of vocalizations. *Ethology*, **119**(10), 831–847. Available from: https://doi.org/10.1111/eth.12122

Schwartz, J.W., Sanchez, M.M. & Gouzoules, H. (2022) Vocal expression of emotional arousal across two call types in young rhesus macaques. *Animal Behaviour*, **190**, 125–138. Available from: https://doi.org/10.1016/j.anbehav.2022.05.017

Selbmann, A., Deecke, V.B., Filatova, O.A., Fedutin, I.D., Miller, P.J.O., Simon, M. et al. (2023) Call type repertoire of killer whales (*Orcinus orca*) in Iceland and its variation across regions. *Marine Mammal Science*, **39**(4), 1136–1160. Available from: https://doi.org/10.1111/mms.13039

Sethi, S.S., Jones, N.S., Fulcher, B.D., Picinali, L., Clink, D.J., Klinck, H. et al. (2020) Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(29), 17049–17055. Available from: https://doi.org/10.1073/pnas.2004702117

Soltis, J., Blowers, T.E. & Savage, A. (2011) Measuring positive and negative affect in the voiced sounds of African elephants (Loxodonta africana). *The Journal of the Acoustical Society of America*, **129**(2), 1059–1066. Available from: https://doi.org/10.1121/1.3531798

Stoeger, A.S. (2021) Elephant sonic and infrasonic sound production, perception, and processing. In: Rosenfeld, C.S. & Hoffmann, F. (Eds.) *Neuroendocrine regulation of animal vocalization*. Academic Press, Elsevier, pp. 189–199. Available from: https://doi.org/10.1016/B978-0-12-815160-0.00023-2

Stoeger, A.S. & de Silva, S. (2014) African and Asian Elephant vocal communication: a cross-species comparison. In: Witzany, G. (Ed.) *Biocommunication of Animals*. Dordrecht: Springer, pp. 21–39. Available from: https://doi.org/10.1007/978-94-007-7414-8_3

Stoeger, A.S., Baotic, A. & Heilmann, G. (2021) Vocal creativity in elephant sound production. *Biology*, **10**(8), 750. Available from: https://doi.org/10.3390/biology10080750

Stoeger, A.S., Heilmann, G., Zeppelzauer, M., Ganswindt, A., Hensman, S. & Charlton, B.D. (2012) Visualizing sound emission of elephant vocalizations: evidence for two rumble production types. *PLoS ONE*, **7**(11), e48907. Available from: https://doi.org/10.1371/journal.pone.0048907

Stoeger, A.S., Zeppelzauer, M. & Baotic, A. (2014) Age-group estimation in free-ranging African elephants based on acoustic cues of low-frequency rumbles. *Bioacoustics – The International Journal of Animal Sound and its Recording*, **23**(3), 231–246. Available from: https://doi.org/10.1080/09524622.2014.888375

Stowell, D. (2022) Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, **10**, e13152. Available from: https://doi.org/10.7717/peerj.13152

Swider, C.R. (2023) Acoustic behavior, poaching risk, and habitat use in African forest elephants (*Loxodonta cyclotis*): insights from passive acoustic monitoring. Ph.D. Syracuse University. Available at: https://www.proquest.com/docview/2827818557/abstract/889382C334AE493CPQ/1 [Accessed 7 August 2024]

Varma, S. & Simon, R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**(1), 91. Available from: https://doi.org/10.1186/1471-2105-7-91

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. et al. (2017) Attention is all you need. Advances in Neural Information Processing Systems [Preprint].

Verahrami, A.K. (2023) Forest elephants modulate their behavior to adapt to sounds of danger. M.S. Colorado State University. Available at: https://www.proquest.com/docview/2852283509/abstract/1465FED374EE472CPQ/1 [Accessed 7 August 2024]

Weiss, K., Khoshgoftaar, T.M. & Wang, D. (2016) A survey of transfer learning. *Journal of Big Data*, **3**(1), 9. Available from: https://doi.org/10.1186/s40537-016-0043-6

Wood, C.M., Klinck, H., Gustafson, M., Keane, J.J., Sawyer, S.C., Gutiérrez, R.J. et al. (2021) Using the ecological significance of animal vocalizations to improve inference in acoustic monitoring programs. *Conservation Biology*, **35**(1), 336–345. Available from: https://doi.org/10.1111/cobi.13516

Wood, J.D., McCowan, B., Langbauer, W.R., Jr., Viljoen, J.J. & Hart, L.A. (2005) Classification of African elephant *Loxodonta africana* rumbles using acoustic parameters and cluster analysis. *Bioacoustics*, **15**(2), 143–161. Available from: https://doi.org/10.1080/09524622.2005.9753544

Wrege, P.H., Rowland, E.D., Keen, S. & Shiu, Y. (2017) Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods in Ecology and Evolution*, **8**(10), 1292–1301. Available from: https://doi.org/10.1111/2041-210X.12730

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014) How transferable are features in deep neural networks? In: *Advances in neural information processing systems*. Montreal, Canada: Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html [Accessed 5 January 2025]

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. UMAP Projections of acoustic features of elephant call-types into 2D space. 2D UMAP projections of the acoustic embeddings for forest elephant roars, rumbles, and trumpets for each CNN (VGGish, Perch, YAMNet, and BirdNET). Each point represents a single vocalisation. Colour indicates call-type. Silhouette scores shown per call-type in legend and overall per model in heading.

Figure S2. UMAP Projections of clustered acoustic features of elephant rumble vocalisations in 2D space. 2D UMAP projections of the acoustic embeddings for forest elephant rumbles for each CNN (VGGish, Perch, YAMNet, and BirdNET). Each point represents a single vocalisation. Colour indicates Affinity Propagation identified cluster. Silhouette scores shown overall per model in heading.

Figure S3. Bar chart of rumbles with biological labels and cluster classification per model. x-axis shows cluster number and y-axis shows percentage proportion of a category label (e.g., 'Infant') distributed across all clusters. One bar chart per CNN model per category. Quadrants represent different categories: (a) age, (b) sex, (c) behaviour, (d) distress.

Table S1. GLM coefficients for each CNN model. Rumble acoustic features as dependent variable. Reference for each predictor category shown in italics across 4 predictor categories. *P*-value significance codes: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.