Acoustics Australia
https://doi.org/10.1007/s40857-023-00298-5

ORIGINAL PAPER

Identifying Distinguishing Acoustic Features in Felid Vocalizations Based on Call Type and Species Classification

Danushka Bandara[1] · Karen Exantus[1] · Cristian Navarro-Martinez[2] · Murray Patterson[3] · Ashley Byun[2]

Received: 4 December 2022 / Accepted: 24 May 2023

Abstract
The cat family Felidae is one of the most successful carnivore lineages today. However, the study of acoustic communication between felids remains a challenge due to the lack of fossils, the limited availability of audio recordings because of their largely solitary and secretive behaviour, and the underdevelopment of computational models and methods needed to address these questions. This study attempts to develop a machine learning-based approach which can be used to identify acoustic features that distinguish felid call types and species from one another through the optimization of classification tasks on these call types and species. A felid call dataset was developed by extracting audio clips from diverse sources. Due to the limited availability of samples, this study focused on the Pantherinae subfamily. The audio clips were manually annotated for call type and species. Time–frequency features were then extracted from the dataset. Finally, several multi-class classification algorithms were applied to the resulting data for classifying species and call types. We found that duration, mean mel spectrogram, frequency range, and amplitude range were among the most distinguishing features for the classifications.

Keywords Acoustic feature extraction · Felid vocalization · Classification · Machine learning

1 Introduction

Acoustic communication plays a critical role in many tetrapod species (Chen and Weins et al. 2020). Such vocalizations can transmit objective information about an individual's internal state as well as information critical for intraspecific interactions such as mate selection and territorial defence [8, 22, 30].

The cat family Felidae represents one of the most successful carnivore lineages today with as many as 40 recognized species organized into eight major lineages [21]. Although most adult felids are largely solitary [10], vocalization still plays a critical role in intraspecific communication [24]. There are 14 major discrete and graded calls documented in Felidae [23, 28]. These calls vary in three major structural domains: loudness (amplitude), time (call duration), and pitch (frequency), though there is also considerable variability in other acoustic features such as harmonic structure. Learning does not seem to play a significant role in felid vocalizations. Rather, the species-specific acoustic structure of these calls is largely genetically determined [24, 25].

While attempts to understand the evolution of vocalizations have been made within Felidae (See [22, 25]), such studies are challenging to conduct due to (1) the lack of fossils that would provide anatomical information about sound production and the acoustic characteristics of ancestral vocalizations, (2) the heterogeneity of audio recordings of vocalizations for Felidae species due in part to their largely solitary and secretive behaviour, and (3) the underdevelopment of computational models and methods.

To address these challenges, our study focuses on three major goals:

1. Develop an annotated felid call dataset by combining audio from diverse sources.
2. Objectively detect key acoustic features that distinguish types of vocalizations and identify species-specific differences.

✉ Danushka Bandara
dbandara@fairfield.edu

[1] Department of Computer Science and Engineering, Fairfield University, Fairfield, CT 06824, USA

[2] Department of Biology, Fairfield University, Fairfield, CT 06824, USA

[3] Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA

Published online: 10 June 2023

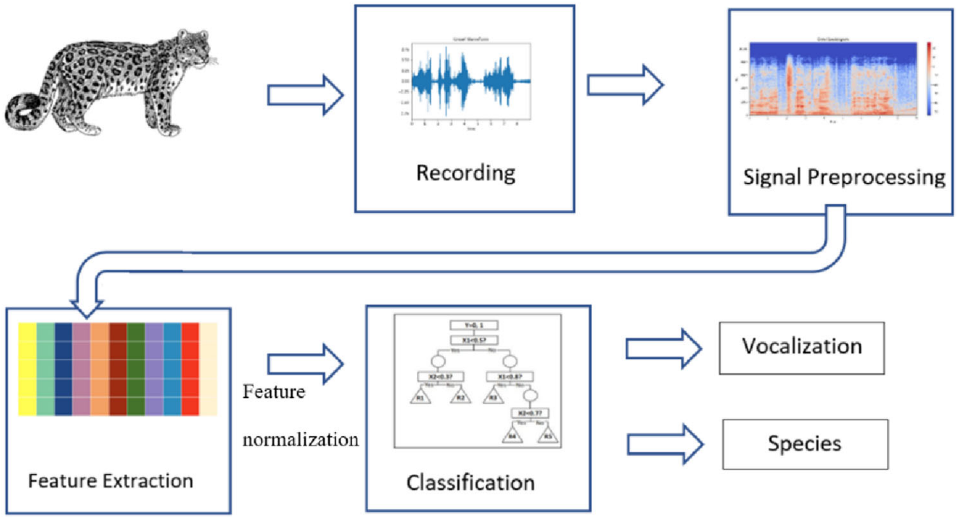


Fig. 1 Flow chart of the machine learning classification of vocalizations and species

3. Develop machine learning classifiers that identifies those features that are distinguishing (to call type or species) through felid call classification from diverse audio sources.

Classification of felid calls has been attempted in several studies, mainly as part of a larger dataset in combination with other animal species calls. For example, Weninger and Schuller [32] developed a classification method by left–right and cyclic Hidden Markov Models (HMMs), recurrent neural networks with Long Short-Term Memory (LSTM), and Support Vector Machines (SVM), achieving up to 81.3% accuracy on a 2-class, and 64.0% on a 5-class classification task. Ji et al. [9] achieved individual animal identification with 90% accuracy using a HMM with frame-based spectral features consisting of cepstral coefficients.

While more recent studies use deep neural networks to achieve felid sound classification (e.g. [19, 20]), the drawback of these methods is their 'black box' nature. These methods, by default, do not provide interpretable acoustic features that can be used for further analysis, such as evolutionary acoustic analysis. Also, these methods require many samples, which is challenging for most felid species.

Due to the greater availability of audio files, in this study we focus specifically on vocalizations from the subfamily Pantherinae which includes species such as lions, tigers, and jaguars. Calls are curated from various sources into a single uniform annotated database standardized for preprocessing and model building. The heterogeneity of audio recordings is addressed by converting the raw audio into the time domain, time–frequency domain, amplitude, cepstral, and statistical features; acoustic features broadly used in previous animal call/classification research (e.g. [1, 3, 12, 14–17, 26]) were used in the analysis. Since these features describe the salient characteristics of auditory signals, they enable combining and comparing raw audio files from diverse sources. The features of the audio clips from each source are normalized separately so that the feature dataset is agnostic of the audio clip source.

We developed a three-class call classification and four-class species classification. Classifiers used amplitude, frequency as well as time-based features. The development of machine learning models provides an automated methodology for gaining insight into the important acoustic features that characterize the different vocalizations in different species from audio without human judgement. Such identification and extraction of features will help promote further research into acoustic communication in felid species.

2 Methods

2.1 Overview

Figure 1 shows the flow chart of the machine learning pipeline. Pantherinae audio clips are collected from online sources, and then manually annotated. Annotated clips are subsequently preprocessed, undergo feature extraction and normalization, and finally input into the classification models, which result in vocalizations and species classification.

2.2 Data Collection

One major challenge this study faces is the lack of a standardized database. The call data used in this study were collected by different researchers for different intended applications. Some other calls were extracted from online videos that were ambient recordings that contain a felid vocalization as part of

Table 1 Sources of felid call audio clips

Calls	No. of audio clips
Museum für Naturkunde	9
YouTube	87
TikTok	22
Panther ridge conservation centre	6
Wildcat conservation centre	1

Table 2 Contents of the felid call dataset by species and call type

Calls	Leopard	Clouded Leopard	Lion	Jaguar
Prusten	0	29	0	11
Roar	29	0	25	11
Growl	12	0	8	0

a longer recording. The audio clips that we use in this study were collected from various sources, including the Animal Sound Archive (Museum für Naturkunde Berlin), calls provided by G. Peters, and social media (see Table 1).

All sound files were manually annotated by domain experts using Raven Pro, GarageBand, and Melodyne 5, which assisted in identifying the specific calls. These manual annotations were saved into individual clips of the complete sound files labelled as the type of felid call. A metadata file was connected with each audio sample, including file name, source, the onset of the call, duration of the call, species name, sex, age, audio quality, and any relevant notes. The distribution of the audio clips by species and call type is provided in Table 2.

Since certain species of cats are both rare and elusive, obtaining a substantial number of audio samples for those species remains a challenge. Thus, the obtained dataset has a class imbalance. As such, the snow leopard species and mew calls were dropped from the subsequent analysis due to an insufficient number of samples for training and test sets.

2.3 Signal Preprocessing

After dataset development, the next step was signal preprocessing. This step involved the conversion of the waveform from the time domain to the frequency (spectrum) and to the time–frequency domain (spectrogram). The spectrograms in Figs. 2 and 3 show that there are acoustic differences between the different call types as well as between species. It can be further seen that both time and frequency domain differences exist between call types and species.

Spectrograms were created using Raven Pro 1.6.4 at a FFT size of 2048, Hann window of 1024 samples and an overlap of 50%. Prusten (or chuffs) are friendly, close-range vocalizations [28]. Significant acoustic differences (t test, $p < 0.05$) noted between each species include duration (s), peak frequency (Hz), peak time (s) and the number of pulses per chuff.

Roaring sequences are considered to function in long distance communication [28]. As shown in Fig. 3, a significant acoustic difference (t test, $p < 0.05$) between leopard and jaguar roars was the fundamental frequency (130 Hz ± 60.30) and (185 Hz ± 56.23), respectively. A greater difference was found between the roaring sequences of the lion and of leopard and jaguar. These include duration (s), peak frequency (Hz), and pulses per second. The lion roar (34.09 s ± 10.58) was significantly longer than both the leopard (7.19 s ± 6.13) or jaguar (8.27 s ± 4.76) and consisted of significantly fewer pulses per second (lion 0.17 ± 0.087, leopard 1.63 ± 0.75, and jaguar 1.72 ± 0.60). The roar between the lion and leopard were additionally distinct in that the lion had a significantly lower peak frequency (lion 218.80 ± 58.16 Hz vs 294.24 ± 164.32) and higher upper frequency range (12,636.93 Hz ± 3710.30 vs 10,107 Hz ± 4855.50).

The growl is a vocalization made by all felids and indicates aggression [28]. Growl spectrograms shown in Fig. 4 show that the growls between these two species were highly similar and only showed a significant difference ($p < 0.05$) in the upper frequency range (leopard 4785.14 Hz ± 2445.28 vs lions 7836.65 Hz ± 3012. 84).

2.4 Acoustic Feature Extraction

Below is a list of features extracted and used for the machine learning models. We used the mean of several of these features in our model, as is common in audio processing [27]. From the literature, we can observe many features being used for animal sound classification. These include: Temporal and spectral features [7], Mel frequency cepstral coefficients [13] and Tonnetz [33]. We have defined below the features that have been selected based on a study of successfully used features from the literature.

- Frequency Range—A continuous range or spectrum of frequencies that extends from one limiting frequency to another.
- Amplitude Range—The difference between the highest positive and the lowest positive amplitude.
- Average Amplitude—The mean of the amplitudes.
- Root-mean-square (RMS) energy—A measure of the audio file's loudness in energy per frame.
- Pulses per Sec—This feature is calculated as the number of peaks per second of the audio clip.
- Duration—The difference between start time and end time from the audio file.

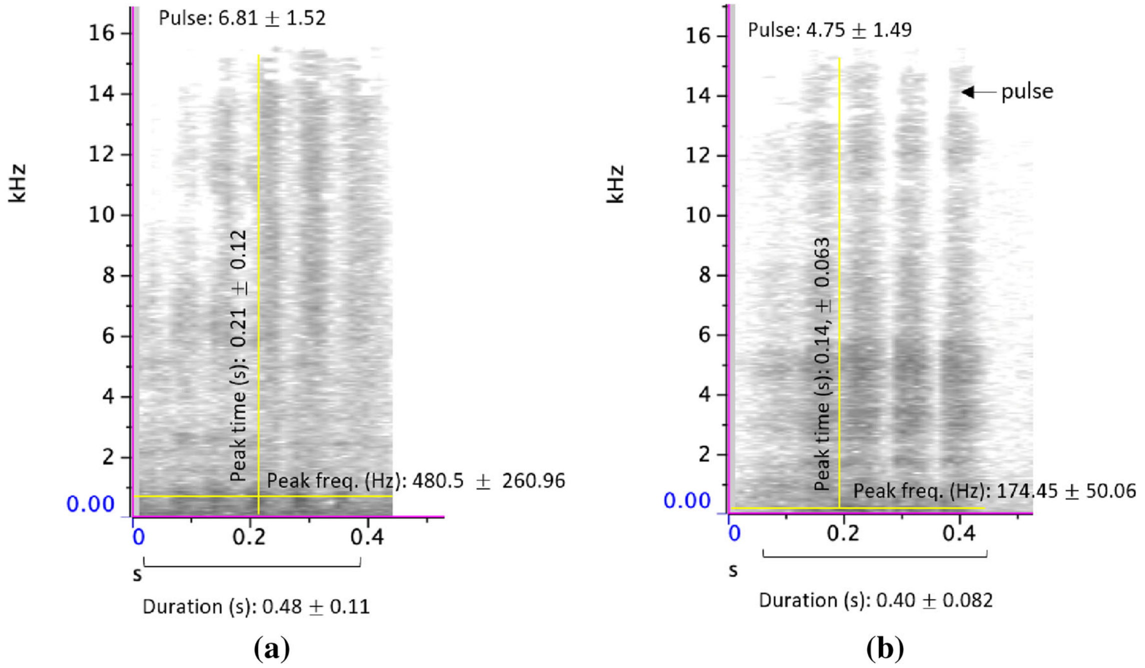


Fig. 2 Spectrograms of **a** clouded leopard and **b** jaguar prusten vocalizations

- Zero Crossings—The number of signal changes between negative and positive values in the spectrum.
- Mel spectrogram—Mel frequency refers to a mapping of audio from the linear frequency domain to the Mel frequency domain. The Mel frequency domain mimics human perception of audio. The Mel spectrogram is the spectrogram of the audio over the Mel scale. This feature consists of the average value of the Mel spectrogram.
- Mel frequency cepstral coefficients (MFCC)—MFCCs are obtained by taking the Mel spectrogram and applying a discrete cosine transformation. The 20 such MFCCs are averaged to obtain this feature.
- Spectral Roll off—This feature computes the roll-off frequency (The spectrogram bin such that at least 85% of the energy of the spectrum in this frame is contained in this bin and the bins below) for each frame in a signal. We used the mean value of the array of frequencies.
- Spectral Contrast—Spectral contrast considers the spectral peak, the spectral valley, and their difference in each frequency sub-band. We used the mean of these values.
- Chromagram—Chromagram is a variation in time–frequency distributions, representing spectral energy. It is computed from a waveform. We used the mean value.
- Tempogram—It is a local autocorrelation of the onset strength envelope. The tempogram computes tempo variation and local pulse in the audio signal. We used the mean value.
- Tonnetz—Represent the lattices of tones. The tonnetz tonal centroids help in Detecting Harmonic Change in Musical Audio or variances due to tones in audio.

The feature statistics for the above described features for each call type and call type per species are shown in Tables 3 and 4, respectively. For call type per species (Table 4), we show only the roar as an example of extracted features.

As discussed earlier, an advantage of our approach is the extraction of acoustic features that can be used for downstream analyses. A preliminary investigation into these features demonstrate significant differences between different call types per species. For example, the roars of jaguars, leopards and lions were distinguishable by a range of extracted features (Table 4). From the 15 features shown, eight show statistically significant differences in feature values between Lion and Jaguar, that number is 13 between lion and leopard. However, we see only three statistically significant differences in features between jaguar and leopard. These values imply that the lion roars were more different from jaguars and leopards than the latter two were to each other. It can also be seen that of the two, the jaguar roar most closely resembled the roar of a lion. We also detected significant differences in the features extracted between species for growls and prustens (data not shown).

2.5 Feature Normalization

The features made up of integers and floating-point values were normalized. By normalizing the features, we were able to address the non-standard nature of the original recordings. Features from each source are normalized to between 0 and 1 before classification. This normalization ensures that each

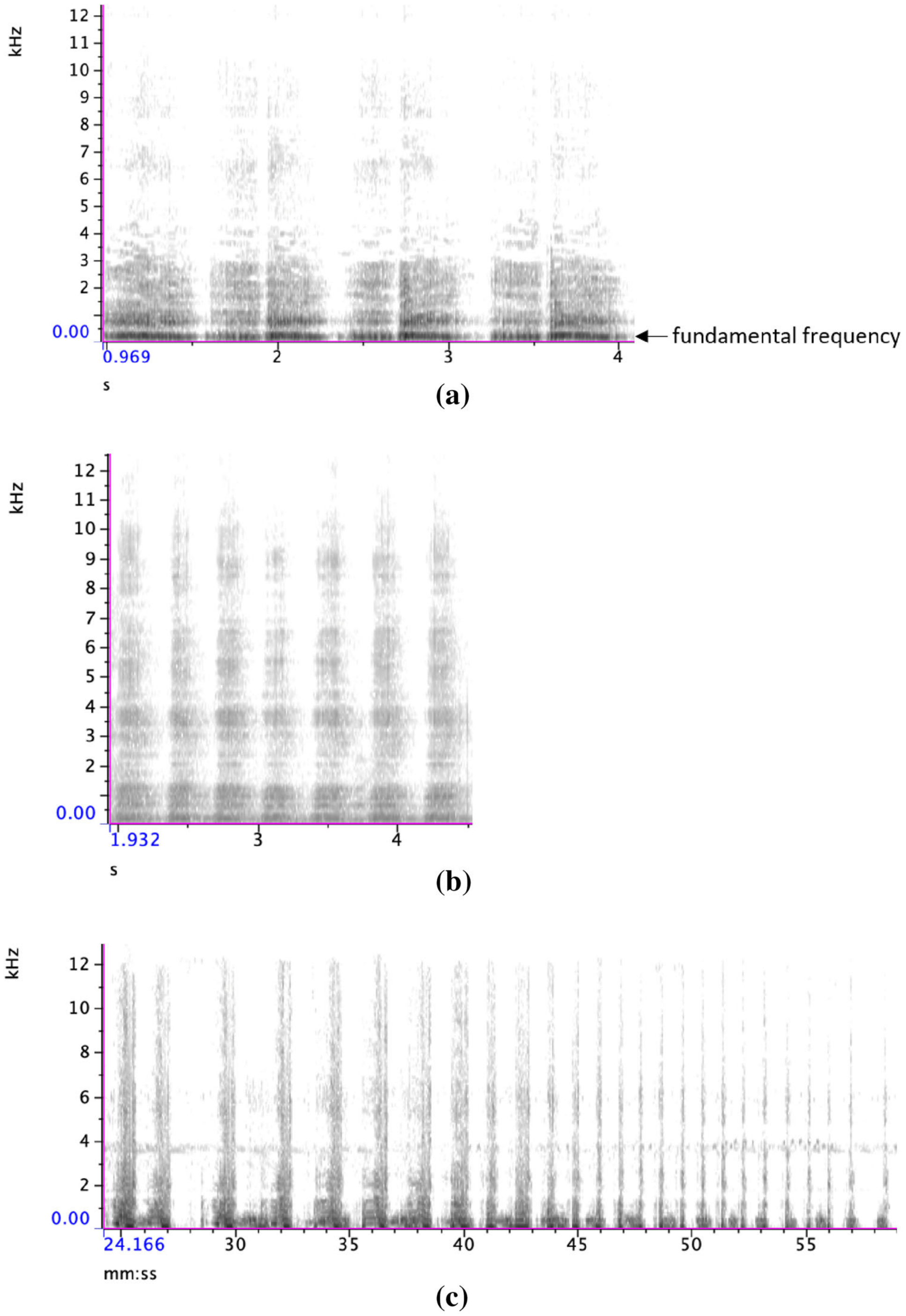
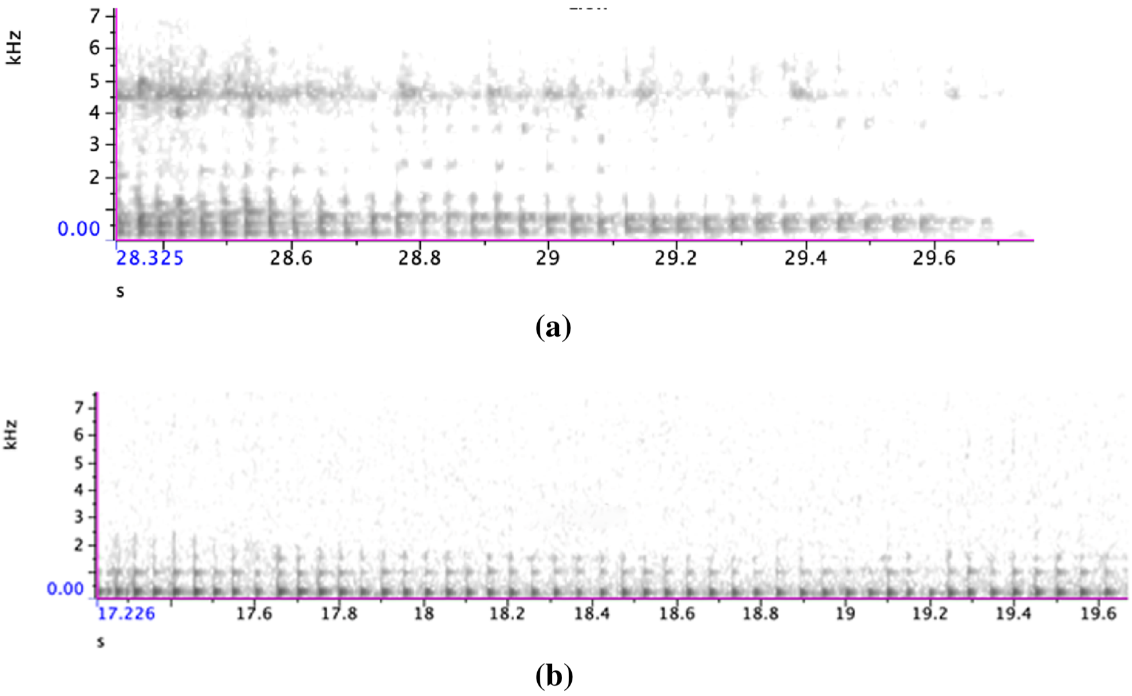


Fig. 3 Spectrograms of **a** leopard, **b** jaguar, and **c** lion roaring sequence vocalizations

source is treated fairly in the classification stage. For example, a feature from an audio clip from YouTube might have a higher amplitude in general than a feature from an audio clip from TikTok. By normalizing, feature magnitude from both sources are rescaled to the same magnitude range. In addition, normalization also helps models that depend on the magnitude of the features (e.g. SVM). This prevents features with higher magnitude from 'overtaking' the loss function. Regularization also benefits from normalization because it prevents certain high-magnitude features from being penalized more than necessary.

2.6 Classification

After data normalization, the features and labels (call or species) progressed to the classification stage. We examined



(a)

(b)

Fig. 4 Spectrograms of **a** lion and **b** leopard growls

the performance of the following five classifiers on call type and species classification:

Support vector machines (SVM): The SVM algorithm attempts to find an optimal hyperplane that separates the training data in Euclidean space. This is achieved by maximizing the margin between two hyperplanes that lie on the support vectors (the data points that belong to two separate classes and are 'closest' to each other).

K nearest neighbours (KNN): This algorithm simply labels test data points based on the labels of the nearest data points to it. The number of nearest data points considered, or K, is a tuneable hyperparameter.

Logistic regression: The logistic regression model learns a logistic function that separates the data points into classes. The model is optimized using gradient descent applied to the loss function (difference between output of model and the actual label).

Random forest: Random forests are formed by multiple decision trees. Decision trees, in turn, are models that split the dataset at multiple levels, choosing the split with the highest 'information gain' at each level.

Gradient boosting: Gradient boosting is a modification of the random forest algorithm such that each tree learns sequentially, i.e. each tree learns the 'residual' from the previous step. For this, we use XGBoost, which is an implementation of gradient boosting.

The evaluation of the classifiers was done using (1) K-fold cross-validation on 70% of the data (training set) and (2) Best model parameters from cross-validation tested on the 30% held-out test set. Cross-validation provides a generalized picture of how the classifier performs over the training set without overfitting to the held-out test set. The number of folds (K) is a hyperparameter that could be optimized. A smaller K would result in smaller training set for the model to train on, while a larger K would be computationally intensive as the number of tests increases with number of folds. In our testing, we found that $K = 10$ provides both a large enough training set in each test, as well as being computationally feasible to perform. First, we split the data in a 70:30 train-test ratio. Then, we applied cross-validation to the 70% training set. In the tenfold cross-validation, 90% of the training set was considered the training data and the rest 10% was considered the validation data. The train-validation cycle was completed 10 times using ten different training and validation splits, and classifier performance was recorded at each cycle, thereby covering the whole training set. The average classification accuracy among the ten folds was considered the cross-validation accuracy for the specific hyperparameters used. After evaluating using cross-validation, the best performing (highest cross-validation accuracy) model parameters were further tested on the held-out 30% test set. The confusion matrix, feature importance, and receiver operating characteristic (ROC) curve were obtained for the held-out test set.

Table 3 Feature statistics for each call type

Sound	Mean	Std. dev	Median	Range
Growl (n = 20)				
Duration (s)	13.40	17.35	8.00	1.00 ↔ 60.0
Zero crossings	26,416.50	29,987.12	17,775.00	612.00 ↔ 114,189.00
Amp range (dB)	0.76	0.22	0.81	0.31 ↔ 1.10
Avg Amp (dB)	0.06	0.02	0.06	0.02 ↔ 0.10
Freq range (Hz)	8894.29	2288.87	10,182.78	3777.00 ↔ 10,658.76
Pulses per sec	1506.39	717.31	1426.77	522.53 ↔ 2740.00
Mean MFCC	− 4.02	2.10	− 3.21	− 8.07 ↔ (− 0.42)
Mean spectral roll off (Hz)	3863.44	1214.67	4064.15	2004.52 ↔ 6174.89
Mean mel spectrogram (dB/Hz)	2.46	1.38	2.41	0.24 ↔ 5.18
Mean spectral contrast (dB)	18.31	1.34	17.90	16.52 ↔ 20.73
Mean chromagram (dB/Hz)	0.71	0.069	0.70	0.56 ↔ 0.82
Mean tempogram (BPM)	0.18	0.067	0.20	0.0098 ↔ 0.27
Mean tonnetz	0.00	0.01	0.00	(− 0.02) ↔ 0.01
Partials (Hz)	49.90	75.16	18.00	4.00 ↔ 330.00
Mean RMS (dB)	0.09	0.03	0.09	0.03 ↔ 0.14
Prusten: Clouded leopard and jaguar (n = 40)				
Duration (s)	1.05	0.22	1.00	1.00 ↔ 2.00
Zero crossings	1810.08	941.27	1446.00	2077.00 ↔ 4053.00
Amp range (dB)	0.19	0.19	0.11	0.03 ↔ 0.92
Avg amp (dB)	0.01	0.0	0.0	0.00 ↔ 0.04
Freq range (Hz)	3310.96	2394.04	2736.5	3194 ↔ 10,681
Pulses per sec	531.64	406.70	470.5	84 ↔ 1941
Mean MFCC	− 7.26	3.27	− 6.40	(− 14.32) ↔ (− 0.976)
Mean spectral roll off (Hz)	3632.48	1248.08	3642.41	1063.93 ↔ 6273.99
Mean mel spectrogram (dB/Hz)	0.25	0.48	0.05	0.003 ↔ 2.32
Mean spectral contrast (dB)	19.86	2.06	20.44	15.91 ↔ 23.23
Mean chromagram (dB/Hz)	0.71	0.09	0.70	0.50 ↔ 0.84
Mean tempogram (BPM)	0.11	0.03	0.11	0.04 ↔ 0.16
Mean tonnetz	− 0.004	0.01	− 0.004	(− 0.03) ↔ 0.02
Partials (Hz)	2.725	3.07	2	0 ↔ 16
Mean RMS (dB)	0.02	0.01	0.01	0.003 ↔ 0.05
Roar (n = 65)				
Duration (s)	17.54	13.73	11	4.0 ↔ 54.0
Zero crossings	30,721.28	30,626.54	16,304	2077 ↔ 127,303
Amp range (dB)	0.64	0.28	0.65	0.05 ↔ 1.04
Avg amp (dB)	0.06	0.06	0.05	0.002 ↔ 0.38
Freq range (Hz)	8335.05	2344.51	9593.43	3194 ↔ 10,681
Pulses per sec	632.76	472.67	468.79	106.75 ↔ 2766.71
Mean MFCC	− 5.83	2.51	− 5.48	(− 10.62) ↔ 1.99
Mean spectral roll off (Hz)	3569.35	1412.56	3627.26	1053.32 ↔ 7885.78
Mean mel spectrogram (dB/Hz)	4.20	7.84	1.91	0.001 ↔ 55.97
Mean spectral contrast (dB)	19.43	1.34	19.13	16.24 ↔ 23.85

Table 3 (continued)

Sound	Mean	Std. dev	Median	Range
Mean chromagram (dB/Hz)	0.66	0.06	0.67	0.44 ↔ 0.78
Mean tempogram (BPM)	0.16	0.05	0.16	0.08 ↔ 0.27
Mean tonnetz	− 0.0006	0.006	− 0.0004	(− 0.01) ↔ 0.02
Partials (Hz)	47.74	46.24	28	0 ↔ 183
Mean RMS (dB)	0.08	0.07	0.06	0.002 ↔ 0.47

3 Results

In the previous section, we described how the dataset of felid calls was created using diverse sources and how the classifiers were subsequently trained and tested on the dataset. In this section, we present the results of the machine learning classification and the most distinguishing features according to the classifier models. Table 5 shows a summary of the best cross-validation accuracies obtained by each classifier.

3.1 Call Classification

The average cross-validation accuracy for the XGBoost model for call classification was 77.37% with a standard deviation of 18 percentage points. Here, we report the detailed analysis of the best performing XGBoost classifier for call classifications retested based on the held-out test set. The confusion matrix (shown in Fig. 5a) is a representation that allows a detailed visualization of a classifier performance. The columns of the matrix show the classes predicted by the classifier for the test data points, while the rows show the actual classes that the data belongs to. The feature importance provides a score that indicates how useful each feature was in the construction of the boosted decision trees within the XGBoost model. The more a feature is used to make key decisions (splits) with decision trees, the higher its feature importance score. The feature importance chart in Fig. 3 is a visualization of the top 13 most important features for the classification. The confusion matrix and most predictive features for this classifier is given in Fig. 5.

The receiver operating characteristic (ROC) curve shows the trade-off between the true positive rate and the false positive rate. Since the ROC curve does not depend on the class distribution, it is useful for evaluating performance in classifications where there is a class imbalance. The ROC curve for the call classification is given in Fig. 6.

The area under the ROC curve (AUC) is a measure of classifier performance. From Fig. 6, we can see that the area under the curve values for all the call types are at or above 0.9.

3.2 Species Classification

The average cross-validation accuracy for XGBoost model for species classification was 76.37% with a standard deviation of 10 percentage points. Here, we report the detailed analysis of the best performing classifier for call classification retested based on the held-out test set. The confusion matrix and feature importance for the classifier are given in Fig. 7a.

From Fig. 8 we can see that the area under the ROC curve for the species classifier is at or above 0.7 for all species. The worst AUC of 0.7 was for the jaguar species. From the most predictive feature analysis, we found that duration, mean mel spectrogram, frequency range and amplitude range appear to be successful in both call and species classifications.

4 Discussion

This study presents an important step towards felid call classification and characterization. In addition to achieving a high degree of accuracy for call type and species, we were also able to extract acoustic features which could be used for additional study on felid vocalizations. Duration, zero crossings, frequency range, amplitude range, mean spectral contrast and mel frequency were shown to be the distinguishing features for both call type and call type per species classifications. We also observed from feature statistics (Tables 3 and 4) that our method allowed us identify significantly different features without any direct human analysis.

Feature extraction for downstream analyses is another advantage of our approach. In the example of the roar, we found significant differences in some of the features between the lion, leopard and jaguar as shown in Table 4. Some of these differences were consistent with what was found using Raven Pro spectrogram analysis such as duration and number of pulses per second. From the features that we extracted, the number of significant feature differences indicated that lions had the most distinct roar of the three felid species compared. This makes sense if we consider that the roar is a long-distance vocalization and that habitat likely influences the acoustic characteristics of this type of call [22]. African

Table 4 Feature statistics for the roar call between jaguar, leopard and lion species

Sound	Mean	Std. dev	Median	Range
Jaguar Roar (n = 11)				
Duration (s)	13.27	7.63	15	4 ↔ 26
Zero crossings	21,143.72	15,751.87	19,758	2077 ↔ 43,089
Amp range (dB)	0.33	0.16	0.31	0.09 ↔ 0.71
Avg amp (dB)	0.03	0.02	0.02	0.01 ↔ 0.07
Freq range (Hz)	7352.94	2780.80	7107.5	3195.88 ↔ 10,434.4
Pulses per sec	647.60	401.75	483.2	168.72 ↔ 1252.31
Mean MFCC	4.79	1.64	− 4.45	− 7.58 ↔ − 2.70
Mean spectral roll off (Hz)	3646.02	1361.64	4150.89	1053.32 ↔ 4899.96
Mean mel spectrogram (dB/Hz)	0.83	1.10	0.49	0.05 ↔ 3.98
Mean spectral contrast (dB)	19.26	1.29	19.44	17.21 ↔ 22.13
Mean chromagram (dB/Hz)	0.67	0.05	0.68	0.54 ↔ 0.72
Mean tempogram (BPM)	0.19	0.06	0.21	0.07 ↔ 0.26
Mean tonnetz	0.002	0.01	0.002	− 0.01 ↔ 0.02
Partials (Hz)	17.63	16.84	18.00	2 ↔ 55
Mean RMS (dB)	0.04	0.02	0.09	0.012 ↔ 0.10
Leopard Roar (n = 29)				
Duration (s)	8.83	6.43	7.00	4 ↔ 39
Zero Crossings	17,258.41	22,431.97	11,439.00	3305 ↔ 122,368
Amp Range (dB)	0.60	0.24	0.55	0.05 ↔ 1.01
Avg Amp (dB)	0.06	0.07	0.04	0.001 ↔ 0.38
Freq range (Hz)	7783.36	2177.76	8052.83	3194 ↔ 10,484.09
Pulses per sec	584.49	552.56	410.20	106.75 ↔ 2766.71
Mean MFCC	− 6.33	3.04	− 6.43	− 10.40 ↔ 1.99
Mean spectral roll off (Hz)	3667.37	1413.30	3631.95	1389.67 ↔ 7885.78
Mean mel spectrogram (dB/Hz)	4.07	10.51	1.48	0.001 ↔ 55.97
Mean spectral contrast (dB)	19.41	1.69	19.07	16.24 ↔ 23.84
Mean chromagram (dB/Hz)	0.67	0.06	0.68	0.44 ↔ 0.78
Mean tempogram (BPM)	0.14	0.04	0.14	0.07 ↔ 0.23
Mean tonnetz	0.00	0.01	0.00	− 0.01 ↔ 0.01
Partials (Hz)	27.62	31.82	18.00	0 ↔ 159
Mean RMS (dB)	0.07	0.09	0.05	0.002 ↔ 0.47
Lion roar (n = 25)				
Duration (s)	29.52	13.43	28.00	4 ↔ 54
Zero crossings	50,552.32	33,765.99	45,704.00	3760 ↔ 127,303
Amp range (dB)	0.82	0.22	0.85	0.24 ↔ 1.03
Avg amp (dB)	0.08	0.04	0.07	0.01 ↔ 0.18
Freq range (Hz)	9407.15	1973.75	10,174.88	3797.5 ↔ 10,681
Pulses per sec	682.22	408.49	483.73	249.67 ↔ 1537.25
Mean MFCC	− 5.72	2.04	− 5.30	− 10.62 ↔ − 2.60
Mean spectral roll off (Hz)	3421.91	1476.86	3379.63	1261.77 ↔ 6503.7
Mean mel spectrogram (dB/Hz)	5.83	5.14	4.87	0.19 ↔ 24.54
Mean spectral contrast (dB)	19.54	0.88	19.42	17.96 ↔ 21.6
Mean chromagram (dB/Hz)	0.65	0.07	0.67	0.49 ↔ 0.74
Mean tempogram (BPM)	0.18	0.04	0.18	0.1 ↔ 0.23

Table 4 (continued)

Sound	Mean	Std. dev	Median	Range
Mean tonnetz	0.00	0.00	0.00	− 0.001 ↔ 0.01
Partials (Hz)	84.32	45.87	87.00	2 ↔ 183
Mean RMS (dB)	0.10	0.04	0.09	0.02 ↔ 0.21

Table 5 Average cross-validation accuracies for call and species identification using various classification algorithms

Model	SVM (%)	kNN (%)	Logistic regression (%)	Random forest (%)	XGBoost (%)
Call classification accuracy	59.55	65.19	58.84	71.73	77.37
Species classification accuracy	55.12	52.75	43.39	71.98	76.73

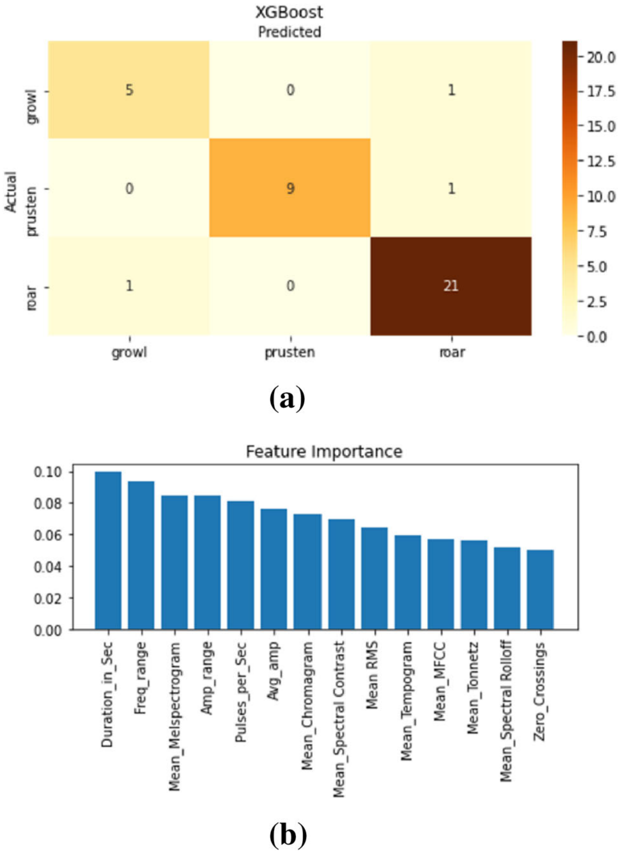


Fig. 5 **a** Confusion matrix and **b** feature importance for the XGBoost call classifier (Accuracy 92%)

lions typically live in dry forests or open plains while jaguar and leopards are usually found in low-lying dense tropical forests [28]. The closer similarity of the leopard and jaguar roar in terms of the features extracted are consistent with these habitat differences illustrating the potential usefulness of these features in future vocalization studies.

Another advantage of our method is the automated and non-biased approach in data collection. Although humans can distinguish between call types and call types per species using spectrograms, an automated method can remove potential biases that come from taking manual measurements and also save time if examining large data sets that could come from long-term vocalization studies.

Currently large data sets for felid vocalizations are not available. Most studies in the literature tackle domestic cats [5, 18, 19, 29] or a single species [31] of cat. This study is one of the first attempts that try to characterize felid calls by building a dataset from heterogenous sources. However, this heterogeneity provided its own set of challenges. Due to the inability to track individual animals in the dataset, we were unable to ensure that there was no data leakage between the training and test sets from the same animal. Nevertheless, due to the heterogeneous nature of the sources, it is quite likely that the audio recordings are from a wide range of individual cats.

Another limitation of this method is that the call types to be detected need to be decided upon a priori, i.e. the method is unable to identify new species or call types without them being in the training set. This is an inherent issue with supervised machine learning methods. To address this limitation, it is possible to use unsupervised methods such as autoencoders or even principle component analysis that project audio calls to a latent space and learn the differences between them in an unsupervised manner.

5 Conclusion

The goal of this research was to develop a standardized dataset of felid calls and identify distinguishing features for felid call and species classification. We showed that amplitude, time and frequency-based features show variation between call types and species types. Then, we developed and validated machine learning classifiers to predict species and call type. The most predictive features were identified and reported for use in future research. The paper also explained the preprocessing, feature engineering, data cleaning, and

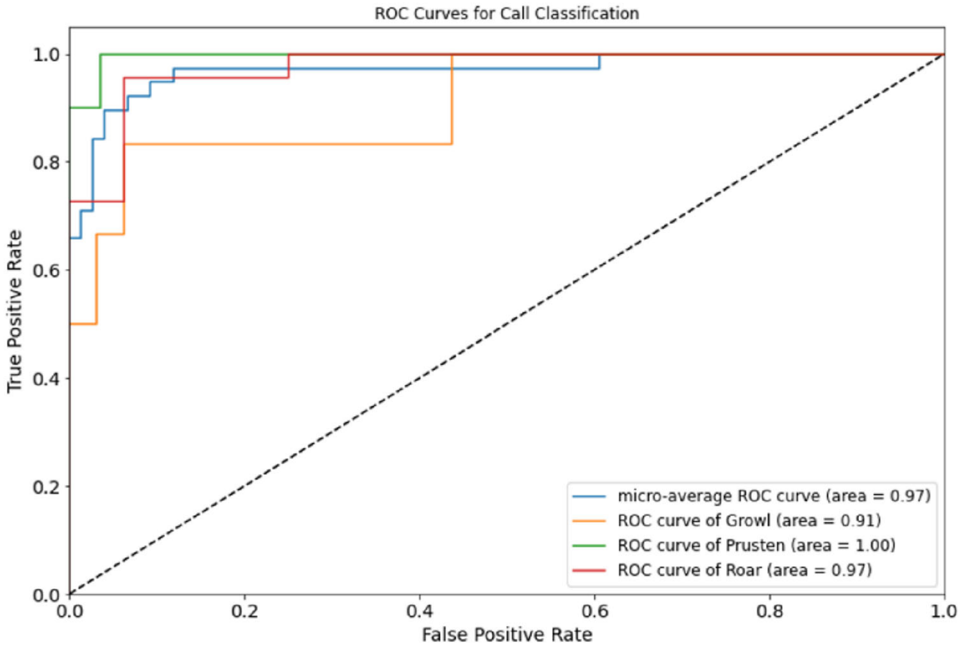
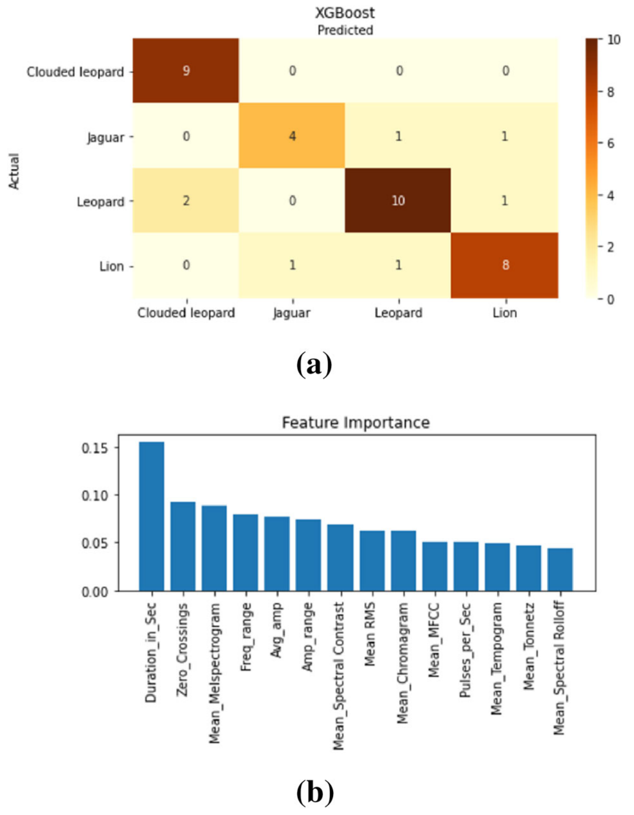


Fig. 6 ROC curves for the XGBoost call classifier



	Clouded leopard	Jaguar	Leopard	Lion
Clouded leopard	9	0	0	0
Jaguar	0	4	1	1
Leopard	2	0	10	1
Lion	0	1	1	8

(a)

(b)

Fig. 7 Confusion matrix (**a**) and feature importance (**b**) for the XGBoost species classifier (82% accuracy)

machine learning steps. Several models were fit to the data, and the best performing was the XGBoost classifier for both call classification and species classification, obtaining 92% and 82% accuracy, respectively. Using cross-validation, we show that on average, we can achieve 77% and 79% accuracy for these two classifications, respectively. Overall, our classifiers were able to obtain better classification performance than comparable multi-class felid call classification studies [32].

These results show that we can classify and characterize the differences in felid calls by vocalization type or species using only an audio recording of their call. Furthermore, significant differences between calls and call types per species among the acoustic features extracted by our model suggest that these data can be useful in further studies on felid vocalizations.

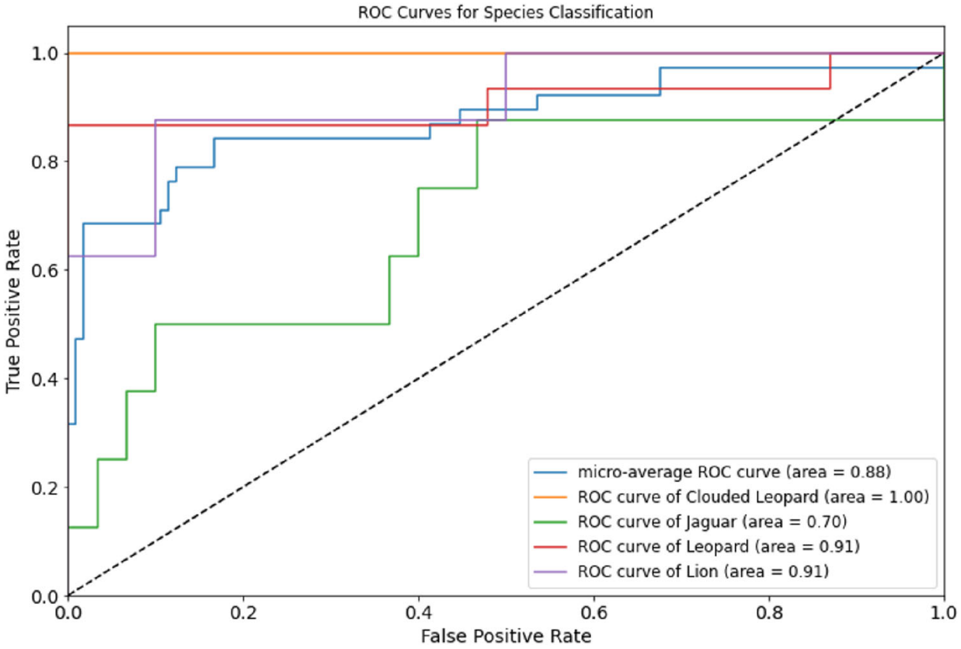


Fig. 8 ROC curves for the XGBoost species classifier

Acknowledgements The authors would like to acknowledge the contributions by: Sai Greeshma Saladi, Rohindraj Kandasamy, Nicholas Furey, Tianyu Yang to the data preprocessing and Brendan Smith, Rebecca Buonopane for assisting in audio annotations. We are grateful for the audio files provided by Dr. Gustav Peters, former Curator of Mammals, Zoological Research Museum, Bonn, and Dr. Karl-Heinz Frommolt, Scientific Head of the Animal Sound Archive (Museum für Naturkunde Berlin). We would also like to thank the anonymous reviewers for their helpful comments that helped to improve this paper.

Funding This work was supported by the Fairfield University Science Institute (AB and MP), the Fredrickson Family Innovation Lab Grant (AB and MP), Fairfield University Computer Science Start-up grant (DB), and Georgia State University Computer Science Start-up Grant (MP).

Data availability The data that support the findings of this study are available on request from the corresponding author, DB. The data are not publicly available due to copyright limitations.

Declarations

Conflict of interest The authors report there are no competing interests to declare.

References

1. Balemarthy, S., Sajjanhar, A., Zheng, J.X.: Our practice of using machine learning to recognize species by voice. arXiv:1810.09078 (2018).
2. Chen, Z., Wiens, J.J.: The origins of acoustic communication in vertebrates. Nat. Commun. **11**, 369 (2020). https://doi.org/10.1038/s41467-020-14356-3
3. Davis, S.B., Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. ASSP **28**, 357–366 (1980)
4. Ehret, G.: Development of Sound communication in mammals. Adv. Study Behav. **11**, 179–225 (1980). https://doi.org/10.1016/S0065-3454(08)60118-7
5. Ferdiana, R., Dicka, W.F., Boediman, A.: Cat sounds classification with convolutional neural network. Int. J. Electric. Eng. Inf. **13**(3), 755–765 (2021)
6. Figueiró, H.V., et al.: Genome-wide signatures of complex introgression and adaptive evolution in the big cats. Sci. Adv. **3**, e1700299 (2017)
7. Gunasekaran, S., Revathy, K.: Content-based classification and retrieval of wild animal sounds using feature selection algorithm. In: 2010 Second International Conference on Machine Learning and Computing, pp. 272–275. IEEE (2010)
8. Hauser, M.D.: The Evolution of Communication. The MIT Press, New York (1996)
9. Ji, A., Johnson, M.T., Walsh, E.J., McGee, J., Armstrong, D.L.: Discrimination of individual tigers (Panthera tigris) from long distance roars. J. Acoust. Soc. Am. **133**(3), 1762–1769 (2013)
10. Johnson et al.: The late miocene radiation of modern felidae: A Genetic Assessment. Science 311: 73–78 (2006). Kitchener, A.: The Natural History of the Wild Cats. A & C Black, London (1991)
11. Kitchener, A.C., Breitenmoser-Würsten, Ch., Eizirik, E., Gentry, A., Werdelin, L., Wilting, A., Yamaguchi, N., Abramov, A.V., Christiansen, P., Driscoll, C., Duckworth, J.W., Johnson, W., Luo, S.-J., Meijaard, E., O'Donoghue, P., Sanderson, J., Seymour, K., Bruford, M., Groves, C., Hoffmann, M., Nowell, K., Timmons, Z., Tobe, S.: A revised taxonomy of the Felidae. The final report of the Cat Classification Task Force of the IUCN/SSC Cat Specialist Group. Cat News Special Issue 11 (2017)
12. Kukushkin, M., Ntalampiras, S.: Automatic acoustic classification of feline sex. In: Audio Mostly 2021, pp. 156–160 (2021)
13. Lee, C.H., Chou, C.H., Han, C.C., Huang, R.Z.: Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Patt. Recogn. Lett. **27**(2), 93–101 (2006)

14. Nanni, L., Maguolo, G., Paci, M.: Data augmentation approaches for improving animal audio classification. Eco. Inform. **57**, 101084 (2020)
15. Nanni, L., Brahnam, S., Lumini, A., Maguolo, G.: Animal sound classification using dissimilarity spaces. Appl. Sci. **10**(23), 8578 (2020)
16. Nanni, L., Maguolo, G., Brahnam, S., Paci, M.: An ensemble of convolutional neural networks for audio classification. Appl. Sci. **11**(13), 5796 (2021). https://doi.org/10.3390/app11135796
17. Ntalampiras, S., Ludovico, L.A., Presti, G., Prato Previde, E., Battini, M., Cannas, S., et al.: Automatic classification of cat vocalizations emitted in different contexts. Animals **9**(8), 543 (2019)
18. Ntalampiras, S., Kosmin, D., Sanchez, J.: Acoustic classification of individual cat vocalizations in evolving environments. In: 2021 44th International Conference on Telecommunications and Signal Processing (TSP), Brno, Czech Republic, pp. 254–258 (2021). https://doi.org/10.1109/TSP52935.2021.9522660
19. Pandeya, Y.R., Lee, J.: Domestic cat sound classification using transfer learning. Int. J. Fuzzy Logic Intell. Syst. **18**(2), 154–160 (2018)
20. Pandeya, Y.R., Kim, D., Lee, J.: Domestic cat sound classification using learned features from deep neural nets. Appl. Sci. **8**(10), 1949 (2018)
21. Pecon-Slattery, et al.: Phylogenetic assessment of introns and sines within the Y chromosome using the cat family felidae as a species tree mol. Biol. Evol. **21**, 22299–22309 (2004)
22. Peters, G., Peters, M.K.: Long-distance call evolution in the Felidae: effects of body weight, habitat, and phylogeny. Biol. J. Lin. Soc. **101**(2), 487–500 (2010). https://doi.org/10.1111/j.1095-8312.2010.01520.x
23. Peters, G.: Vocal communication in cats. In: Great Cats, pp. 76–77 (1991)
24. Peters, G.: Vergleichende Untersuchung zur Lautgebung einiger Feliden (Mammalia, Felidae), Vol. 1 of Spixiana, pp. 1–283 (1978)
25. Peters, G., Tonkin-Leyhausen, B.A.: Evolution of acoustic communication signals of mammals: friendly close range vocalizations in Felidae (Carnivora). J. Mamm. Evol. **6**(2), 129–159 (1999)
26. Raccagni, W., Ntalampiras, S.: Acoustic classification of cat breed based on time and frequency domain features. In: 2021 30th Conference of Open Innovations Association FRUCT, pp. 184–189. IEEE (2021)
27. Rana, D., Jain, A.: Effect of windowing on the calculation of MFCC statistical parameter for different gender in hindi speech. Int. J. Comput. Appl. **98**(8), 6–10 (2014). https://doi.org/10.5120/17201-7409
28. Sunquist, M., Sunquist, F.: Wild Cats of the World. University of Chicago Press, Chicago (2002) ISBN 0-226-77999-8
29. Suzuki, Y., Osawa, A.: Identifying individual cats by their chewing sounds using deep learning. In: Stephanidis, C., Antona, M., Ntoa, S. (eds) HCI International 2021—Posters. HCII 2021. Communications in Computer and Information Science, Vol. 1420. Springer, Cham (2021). . https://doi.org/10.1007/978-3-030-78642-7_74
30. Tavernier, C., Ahmed, S., Houpt, K.A., Yeon, S.C.: Feline vocal communication. J Vet Sci. **21**(1), 18 (2020). https://doi.org/10.4142/jvs.2020.21.e18
31. Trapanotto, M., Nanni, L., Brahnam, S., Guo, X.: Convolutional neural networks for the identification of african lions from individual vocalizations. J. Imaging **8**(4), 96 (2022). https://doi.org/10.3390/jimaging8040096
32. Weninger, F., Schuller, B. (2011). Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 337–340. IEEE.
33. Wu, X., Zhou, S., Chen, M., Zhao, Y., Wang, Y., Zhao, X., et al.: Combined spectral and speech features for pig speech recognition. PLoS ONE **17**(12), e0276778 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.