Highlights

One-Class Bioacoustic Detector for Monitoring the Critically Endangered Pied Tamarin (*Saguinus bicolor*)

Juan G. Colonna, Tainara V. Sobroza, Marcelo Gordo, Eduardo F. Nakamura, Alejandro C. Frery

- **Open-set bioacoustic detector for the pied tamarin.** We introduce an open-set pipeline—band-pass (5–10 kHz) → Perch embeddings → one-OCSVM—tailored to *S. bicolor*, filling a gap with no prior species-specific detector.

- **Bird-trained Perch embeddings transfer to primates.** We show that Perch embeddings trained on birds generalize to *S. bicolor* vocalizations, enabling cross-taxon reuse without retraining.

- **Band-pass filtering reduces background false positives and improves separability.** On our evaluation sets, the 5–10 kHz filter removed all background FPs and shifted ROC curves up/left, increasing AUC.

- **Robust cross-site detection on long recordings with very few false positives.** In a ∼10 min Mindú Park test, the detector prioritized precision, with an observed false-positive rate of 0.03, and improved AUC from 0.74 (raw) to 0.83 (filtered).

- **End-to-end monitoring via occupancy modeling.** We couple detections with a single-site occupancy estimator and derive a closed-form MLE to estimate the cross-site false-detection probability.

One-Class Bioacoustic Detector for Monitoring the Critically Endangered Pied Tamarin (*Saguinus bicolor*) 🐒

Juan G. Colonna[a,d], Tainara V. Sobroza[b], Marcelo Gordo[c], Eduardo F. Nakamura[a], Alejandro C. Frery[d]

[a]*Instituto de Computação (IComp), Universidade Federal do Amazonas (UFAM), Manaus, 69080-900, Amazonas, Brasil*
[b]*Universidade Federal do Oeste do Pará (UFOPA), Campus Oriximiná, Oriximiná, 68270-000, Pará, Brasil*
[c]*Departamento de Biologia, Universidade Federal do Amazonas (UFAM), Manaus, 69080-900, Amazonas, Brasil*
[d]*School of Mathematics and Statistics, Victoria University of Wellington (VUW), Wellington, 6012, New Zealand*

Abstract

The pied tamarin (*Saguinus bicolor*) is a critically endangered primate with a small geographic range that includes fragmented urban forest mosaics in Amazonia, where habitat subdivision and anthropogenic actions complicate its survival and monitoring. Passive acoustic monitoring (PAM) offers a convenient, noninvasive way to track this species, yet open-set rainforest soundscapes make single-species detection challenging. We present a machine-learning pipeline with a very low false-positive rate, appropriate for downstream inference. The method combines a band-pass filter (5–10 kHz), Perch bioacoustic embeddings (deep learning), and a One-Class SVM (OCSVM) applied to sliding windows of continuous audio recordings to detect *S. bicolor* calls. We train on a reduced dataset of labeled calls and validate against diverse out-of-class audio (birds, anurans, anthropophony, and geophony/insects), then test on long, cross-site recordings. The approach achieves high discrimination on held-out negatives and produces very low false-positive rates in realistic, continuous audio. Finally, we pair detections with a single-site occupancy model in a cross-site setting to illustrate end-to-end utility for conservation monitoring and to estimate the false-negative detection probability in recordings from pied tamarin populations in a different geographic region. Our strategy provides a tool for PAM of *S. bicolor* that requires minimal manual labeling effort and can be adapted to other open-set, single-species monitoring scenarios. We also release a convenient Python package (`sauim-detector`), installable via `pip`, that processes an audio file and produces detection timestamps as an Audacity label file (.txt), enabling faster manual verification.

Keywords: Passive acoustic monitoring (PAM), Bioacoustics classification, Open-set species detection, One-class SVM (OCSVM), *Saguinus bicolor* (pied tamarin)

1. Introduction

The pied tamarin (*Saguinus bicolor*, Figure 1a) is a small primate of the Brazilian Amazon. This species is endemic to a geographically restricted region of approximately 8,354 km^2 (Lagroteria et al., 2024; Albernaz et al., 2026). Unfortunately, this geographic range includes the city of Manaus, the second-largest urban agglomeration in the Brazilian Amazon. Due to its sensitivity to habitat loss and fragmentation, and its importance to the Amazonian environment through seed dispersal (Fernandes et al., 2024), the species has been monitored for more than four decades (Ayres et al., 1982).

The *S. bicolor* is listed on the IUCN Red List of Threatened Species as a critically endangered species, with its assessment updated in 2021 (Gordo et al., 2021). Within the city of Manaus, several urban forest protected areas (APAs) have served as refuges for the pied tamarin; however, because these are fragmented urban forests, survival remains challenging (Vidal and Cintra, 2006). Vehicle collisions, electrocution on power lines, and habitat degradation due to human encroachment are among the main threats (Gordo et al., 2013). In 2011 the Brazilian government created a National Plan for the conservation of the *Sauim-de-coleira*. Initiatives have been coordinated to study and

Email addresses: `juancolonna@icomp.ufam.edu.br` (Juan G. Colonna), `tv.sobroza@gmail.com` (Tainara V. Sobroza), `mgordo@ufam.edu.br` (Marcelo Gordo), `nakamura@icomp.ufam.edu.br` (Eduardo F. Nakamura), `alejandro.frery@vuw.ac.nz` (Alejandro C. Frery)

(a) Two individuals of pied tamarin *S. bicolor*.

(b) Map of the central Brazilian Amazon showing the range of the pied tamarin (*Saguinus bicolor*) outlined in orange; the Amazon and Negro rivers are labeled, and Manaus is marked with a red dot. `https://shorturl.at/8AUTv`

Figure 1: (a) Pied tamarins (*S. bicolor*) photographed in an Environmental Protection Area (APA) on the Federal University of Amazonas campus. (b) The species' geographic range overlapping with the city of Manaus in the Brazilian Amazon. Adapted from Lagroteria et al. (2024); Albernaz et al. (2026).

monitor populations to support evidence-based management with the aim of safeguarding the species (ICMBio, 2011).

Among the several available ways to detect the presence of the *S. bicolor*, the passive acoustic monitoring (PAM) systems are among the least intrusive methods (Sobroza et al., 2025). With PAM, autonomous acoustic recorders capture the Amazonian soundscape, and the presence of the pied tamarin can be subsequently inferred from its vocalizations (calls). Recorders offer wide detection ranges and $\approx 360^\circ$ coverage without requiring line of sight — as would a camera trap (Crunchant et al., 2020). However, the abundance of non-target sounds and the dynamic forest background impose a substantial challenge for reliable acoustic detection.

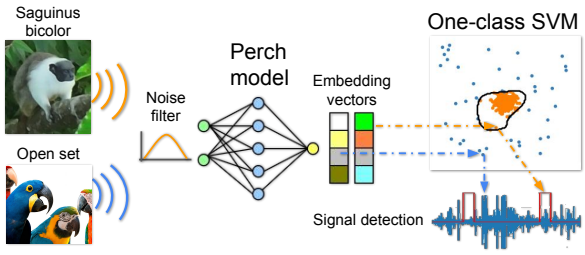
High-quality acoustic classifiers for other animal groups have been proposed (e.g., BirdNET for birds (Kahl et al., 2021) and howler monkeys (Wood et al., 2023); anuran call detectors (Colonna et al., 2016)), but none specifically target *S. bicolor*. Training a multi-class—or even a binary—classifier for a single species can be tricky: what should constitute the negative class? The Amazonian soundscape is rich, including birds, anurans, other mammals, insects, geophonies (e.g., wind, rain, rivers) (Xavier et al., 2024), and anthropophonies (e.g., traffic, voices, machinery) due to proximity to urban areas or human incursions (Sobroza et al., 2024). This diversity—and the fact that the set of non-target sounds is effectively unbounded and shifts across sites and seasons—characterizes an open-set recognition problem (Geng et al., 2021). Consequently, a one-class or anomaly-detection approach is appropriate for *S. bicolor* detection, as we adopt in this work (Figure 2a).

Fortunately, the pied tamarin has a distinctive, nearly unique call signature, except in zones of sympatry where golden-handed tamarins (*S. midas*) may shift their calls toward the pied tamarin's pattern (Sobroza et al., 2021). The characteristics of *S. bicolor* vocalizations can be seen in the spectrogram in Figure 2b. A spectrogram is a two-dimensional representation of an acoustic signal, with time on the horizontal axis, frequency on the vertical axis, and energy encoded as color intensity. This representation enables detailed analysis of the temporal and spectral structure of animal vocalizations and is essential in acoustic studies.

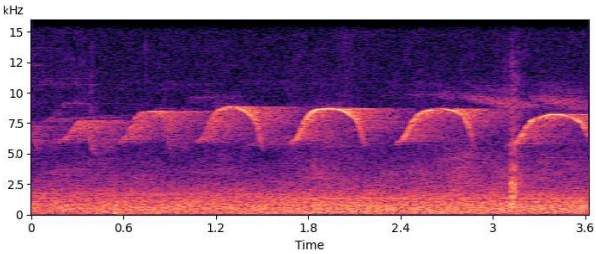
In *S. bicolor*, call energy is concentrated predominantly between approximately 6 and 9 kHz. The calls consist of a sequence of arch-shaped patterns (i.e., syllables) repeated over the course of the signal, typically comprising 1–6 syllables with a total duration of roughly 2–5 s (Sobroza et al., 2017). Such variation may reflect different behavioral contexts, including group cohesion, territory defense, or even adjustments to anthropogenic sounds (Sobroza et al., 2024). This domain knowledge is incorporated into our method via a band-pass filter that targets this frequency band (Section 3.5).

The distinctive and consistent acoustic features of pied tamarins (e.g., call duration, syllable repetition rate, and characteristic pattern shapes) provide reliable species-specific signatures that can be leveraged in passive acoustic monitoring (PAM) systems, enabling non-invasive detection and long-term population tracking. Because our objective is to autonomously and specifically monitor a single species of interest, an open-set recognition formulation is required to handle signals from previously unseen classes (Section 2).

In the present study, *S. bicolor* vocalizations constitute the sole "known" class, and a One-Class Support Vector Machine (OCSVM) is applied to detect calls consistent with this class while rejecting sounds from other species and background noise. Conceptually, one-class classification can be viewed as a special case of



(a) Perch-based embedding extraction.



(b) Example of a spectrogram of a pied tamarin call.

Figure 2: (a) Conceptual overview of the proposed approach. (b) A spectrogram showing a typical call of this species.

open-set classification, in which only a single target class is modeled during training (Khan and Madden, 2014). We propose a method that an incoming audio recording is first normalized to ±1 amplitude, band-pass filtered (e.g., 5–10 kHz), and segmented into five-second windows. Then each segment is passed through the Perch bioacoustic embedding model to obtain an embedding vector. Finally, these vectors are classified by an OCSVM trained only on pied tamarin calls (Figure 2a).

While open-set classification assigns inputs to one of several known classes or rejects them as unknown, one-class classification defines a boundary that encloses only the target class in feature space (e.g. Figure 2a), treating all other inputs as unknown or anomalous (Marques et al., 2023). Therefore, our approach addresses the open-domain challenge of bioacoustic monitoring in complex soundscapes, where signals from unobserved classes are inevitably present.

2. Problem formulation

An open-set problem is a setting in which, at test time (deployment), classes not seen during training may appear (Scheirer et al., 2013). The model must both recognize the known classes and be able to reject samples from unknown classes (rather than forcing a misclassification into a known class). Formally, let K be the set of classes available during training and C the set of classes that may appear at test time, with $K \subseteq C$. The unknown classes are $U = C \setminus K$. An open-set classifier implements a decision rule $f : \mathcal{X} \to K \cup \{\text{unknown}\}$ that (i) assigns correct labels to inputs $x \in \mathcal{X}$ from classes in K and (ii) rejects inputs from U, avoiding high-confidence predictions in regions of the feature space without training support. In our case, $|K| = 1$; therefore, we adopt a one-class model based on the One-Class Support Vector Machine (Schölkopf et al., 2001).

In this study, each *Saguinus bicolor* vocalization is represented by a 1280-dimensional embedding vector extracted from a 5-second audio segment using Google's Perch bioacoustic model (Google Research, 2025). Formally, the feature extractor can be seen as a mapping $f : X \to V, \quad x \mapsto \mathbf{v} \in \mathbb{R}^{1280}$, where X denotes the space of 5-second acoustic waveforms and V the resulting feature space. We hypothesize that these embedding vectors capture salient spectral—temporal patterns of the calls, enabling robust separation of *S. bicolor* vocalizations from other all other sounds in a complex and diverse rainforest soundscape.

To model the target class, we apply OCSVM, which seeks to find a decision boundary in the high-dimensional embedding space that encloses most of the training vectors while allowing a small fraction of them to fall outside (Schölkopf et al., 1999). In its primal form, the OCSVM solves:

$$\min_{\mathbf{w},\rho,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho, \tag{1}$$

subject to $(\mathbf{w} \cdot \mathbf{v}_i) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n,$ where:

- $\mathbf{v}_i \in \mathbb{R}^{1280}$ is the embedding vector of the i-th vocalization segment;
- $\mathbf{w}$ is the normal vector defining the separating hyperplane in the feature space;
- ρ is the offset defining the decision threshold;
- ξ_i are slack variables that permit some training points to be considered outliers;
- n is the number of training examples;
- $\nu \in (0, 1]$ controls the maximum fraction of training outliers and the minimum fraction of support vectors.

The decision function for a new embedding vector $\mathbf{v}$ is $f(\mathbf{v}) = \text{sign}\,(\mathbf{w} \cdot \mathbf{v} - \rho)$, returning $+1$ if the segment is classified as a *S. bicolor* vocalization and -1 if classified as an anomaly (i.e., another species, environmental sound, anthropogenic noise, etc).

In the OCSVM formulation (Eq. 1), the hyperparameter $\nu \in (0, 1]$ plays a dual role. First, ν sets an upper bound on the fraction of training samples allowed to lie outside the learned decision boundary. Second, it sets a lower bound on the fraction of support vectors and, together with the kernel parameters (e.g., γ for the RBF kernel), governs the complexity of the decision function. In the context of *S. bicolor* acoustic monitoring, setting ν too high yields a very tight boundary that may overfit the variability of training calls, misclassifying legitimate vocalizations as anomalies. Conversely, setting ν too low produces a looser boundary that risks admitting non-target sounds—other animal vocalizations, geophonic events (e.g., wind, rain), or anthrophonic noise—as *S. bicolor* calls.

3. Materials and method

We used five datasets in the experiments that follow. The main dataset contains recordings of *S. bicolor* vocalizations captured outdoors under real Amazonian rainforest conditions. The recordings were made using a Sennheiser ME67 unidirectional microphone matched to a Zoon H4N recorder. The remaining four datasets comprise bird and anuran calls, as well as anthropophony, geophony, and insects noise. This diversity of sounds is intended to cover scenarios likely to occur during field monitoring with acoustic sensors.

3.1. S. bicolor and background noises dataset

The *S. bicolor* dataset consists of two sets of recordings: 32 recordings from Bosque da Ciência at the Instituto Nacional de Pesquisas da Amazônia (INPA) and 38 recordings from Mindú Park. These sets come from two different groups of individuals and exhibit distinct background noise conditions. Having multiple specimens is desirable for testing the transferability and generalization of the trained model, as calls may share global acoustic characteristics while also showing subtle, individual-specific differences.

From the first group we extracted 118 five-second segments containing pied tamarin vocalizations. This set was split (holdout) into 82 calls for training and 36 for validation. We also extracted 179 segments containing only background noise. These 179 noise segments were combined with the 36 held-out call segments to form a 215-sample binary test set. Thus, the OCSVM was trained on 82 *S. bicolor* calls and evaluated on 215 test samples. Each sample was labeled by the OCSVM as an inlier (target) or outlier (non-target). We refer to this setup as the "Background" experiment in our figures and tables.

3.2. Birds and Anurans datasets

Common sources of bioacoustic events in rainforests are vocalizations produced by birds and anurans (frogs and toads). Many bird species vocalize in high-frequency bands, increasing the likelihood of overlap with *S. bicolor* calls and, consequently, leading to misclassification. In this study, we curated 30 recordings from 26 bird species obtained from the collaborative Xeno-canto repository[1], totaling 130 five-second record segments. We selected only recordings from the city of Manaus and its surroundings, overlapping the geographic range of the *S. bicolor* individuals used to train our OCSVM.

The anuran dataset includes 60 recordings from ten species, some of which occur frequently in the same areas as S. bicolor, while others are more broadly distributed across the Amazon. In total, we obtained 122 five-second acoustic segments. Note that most of these amphibians, which could potentially confuse the machine learning algorithm, vocalize at night, whereas the pied tamarin vocalizes during the day. Complete lists of bird and anuran species are provided in Section Appendix A. We refer to these two datasets as "Birds" and "Anurans"; they serve as outlier sets to evaluate our method's performance on acoustic events outside the target class.

3.3. Anthropophony dataset

The anthropophony (human-generated sounds) dataset was assembled from the ESC-50 corpus (Piczak, 2015). We concatenated variable-length segments into a single long recording comprising three airplane segments, one car horn, three city traffic segments (including horns), three chainsaw segments, two church bells, one applause segment, three coughs, four miscellaneous engine sounds (e.g., tractor, motor), three fireworks, three hand saw, two helicopter, three train, and three different siren recordings.

These sources reflect plausible human activities near protected areas—for instance, fireworks during soccer match days can be audible across urban reserves; engines and motors are common at reserve edges; chainsaw and hand saw sounds may indicate illegal logging; and helicopters occasionally overfly monitored sites. Although some of these sounds (e.g., bells or trains) are not typical of the areas occupied by the pied tamarin, these negative samples help assess our method's capability to generalize to other scenarios. We standardized these clips into 84 five-second segments, which we use as the "Anthropophony" set.

[1] `https://xeno-canto.org/`

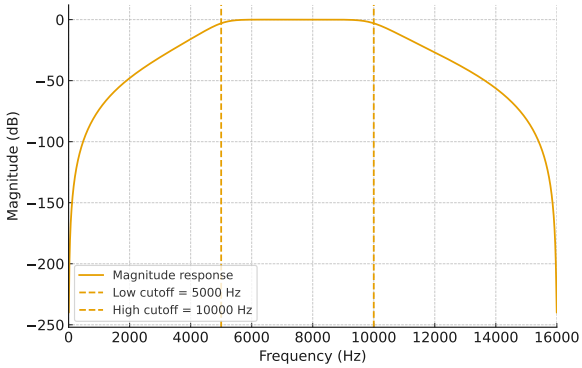


Figure 3: Butterworth band-pass filter.

3.4. Geophony and insect sounds

The geophony and insects dataset includes three crackling fire segments, three cricket segments, four other insect recordings (mostly flies), five rain segments, three sea waves, four thunderstorms, two water drops, and five wind segments. Although some of these sounds—for example, sea waves—are not common in Amazonia, they add diversity to the negative class in this experiment. Because our strictly biophonic material was limited, we grouped common insect sounds (e.g., flies, crickets, bees) with geophony for this class, as both can introduce broadband or quasi-stationary energy that may mask target calls. After standardization, this collection totals 77 five-second segments, which we refer to as the "Geophony" set.

3.5. Band-pass noise filter

Figure 2b shows that the energy of *S. bicolor* vocalizations is concentrated in the 5–10 kHz band, whereas rainforest sounds lie outside this range or occasionally overlap it. For example, in the 3.1–3.6 s interval this spectrogram exhibits transient events that partially overlap the calls. Such overlapping interference degrades the quality of features extracted by the Perch model and can reduce detection/recognition performance. To mitigate this, we include a Butterworth band-pass filter centered on the target band as the first stage of our classification pipeline. Butterworth low- or high-pass filters are widely used in bioacoustics for their simple parameterization and smooth, monotonic magnitude response (Miksis-Olds and Tyack, 2009; Laidre et al., 2024).

We applied a fourth-order ($N = 4$) IIR Butterworth band-pass filter with cutoff frequencies $f_c \in [5, 10]$ kHz, implemented as cascaded second-order sections (SOS). The Butterworth response is maximally flat in the passband and monotonic across the spectrum, with an asymptotic roll-off of $\approx 6N = 24$ dB/oct on each side. Figure 3 shows the magnitude response (dB) for a sampling rate $f_s = 32$ kHz. Therefore, our recorded signals were processed with zero-phase forward–backward filtering using `sosfiltfilt`[2] from the SciPy Python library, which eliminates phase distortion while preserving amplitude characteristics. The effect of this filter on *S. bicolor* vocalizations is discussed later (see Section 6).

3.6. Perch: bioacoustic embedding model

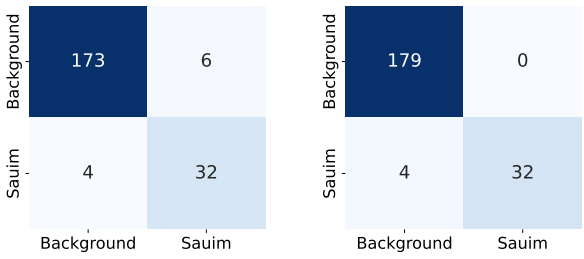
We used embeddings from the Perch bioacoustic model available in the Kaggle public repository (Google Research, 2025), which is based on an EfficientNet-B1 convolutional neural network architecture with approximately 7.8M parameters trained on a global corpus of bird vocalizations (Xeno-Canto) with more than 15k species (Ghani et al., 2023). This model expects input as single-channel five-second audio segments sampled at 32 kHz; it outputs logits and an embedding vector for each segment.

The Perch model has been compared extensively with BirdNET, AudioMAE, YAMNet, and VGGish, showing strong transfer-learning performance across diverse bioacoustic tasks (Ghani et al., 2023). Derivative models such as SurfPerch further demonstrate adaptability to non-avian domains (e.g., tropical reefs) (Williams et al., 2025). Importantly, Perch embeddings are highly transferable and data-efficient, enabling few-shot training and an active-learning loop that rapidly surfaces additional positive samples via vector search, yielding high-quality detectors with modest annotation effort (Dumoulin et al., 2025). These properties make Perch an ideal choice for our approach given our small annotated dataset. Here, we further show that Perch transfers to primates by training a one-class SVM on its embeddings to perform open-set detection.

4. Experimental setup

In this study, we adjust ν (Eq. 1) by grid search over $[0.01, 0.10]$, choosing the value that maximized area under the ROC curve (AUC) on a held-out validation set. This criterion yielded a balance between sensitivity to target vocalizations and robustness to the heterogeneous Amazonian soundscape. The kernel function was chosen as the radial basis function (RBF), which is well-suited for modeling non-linear class boundaries in high-

[2] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.sosfiltfilt.html`



(a) Without filtering. (b) Band-pass filtering.

Figure 4: Confusion matrices comparing call detection before and after band-pass filtering on the validation set ($n = 36$). Band-pass filtering eliminates false positives ($6 \rightarrow 0$). Rows correspond to true labels; columns to predicted labels.

dimensional spaces, and the γ hyperparameter was automatically set according to the rule $1/\text{features}$, corresponding to the inverse of the 1280-dimensional embedding size. The OCSVM implementation was provided by the `scikit-learn` Python library (Pedregosa et al., 2011).

All codes to reproduce our results are provided as ready-to-run, interactive Jupyter notebooks, together with the data used in our experiments, at `https://github.com/juancolonna/Sauim/tree/main`.

5. Results

Our experimental results are divided into three main sections. The first section addresses the classification problem using the five datasets described in Section 3. The second addresses detection in long recordings (Sec. 6). Section 7 shows how a bioacoustic classifier can be coupled with an occupancy model to estimate true- and false-positive detection probabilities (p_{11} and p_{10}) conditional on site occupancy. We conducted classification and detection experiments with and without a band-pass filter to assess its impact on the method's performance.

5.1. Classification

Our method comprises a five-stage pipeline: (i) load each recording and resample to the Perch input rate ($f_s = 32\,\text{kHz}$); (ii) apply the Butterworth band-pass filter; (iii) segment the audio with a five-second sliding window; (iv) pass all segments through the Perch neural network to obtain embedding (feature) vectors; and (v) train the OCSVM on the resulting embeddings.

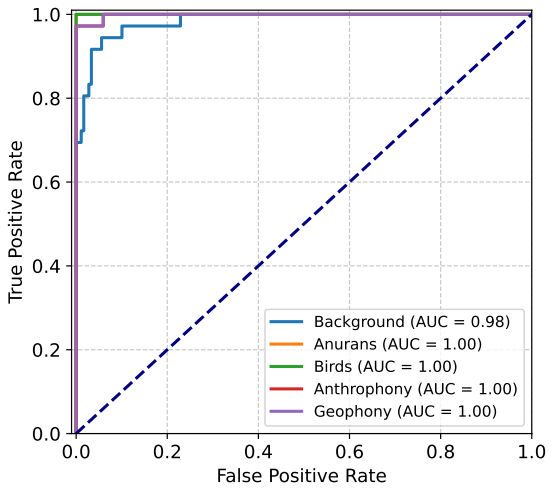
Of the 118 five-second *S. bicolor* call segments, 82 were used to fit the OCSVM (positive-only training). The remaining 36 call segments were combined with 179 background-noise segments to form the test set. Figure 4 compares confusion matrices obtained with and without band-pass filtering prior to feature extraction.

The OCSVM outputs a decision score s, with $s > 0$ indicating inliers (call segments) and $s < 0$ indicating outliers (non-target). Segments lying on the soft-margin decision boundary are assigned the label 0. We adopt the rule $\hat{y} = \mathbf{1}[\, s \geq 0\,]$, i.e., classify as inlier when $s \geq 0$. The band-pass filter eliminates the six false positives that appear when the recordings are not filtered (Figure 4b). Additional confusion matrices for the remaining negative classes (anurans, birds, anthropophony, and geophony+insects) are shown in Figure B.10 provided in Appendix B.

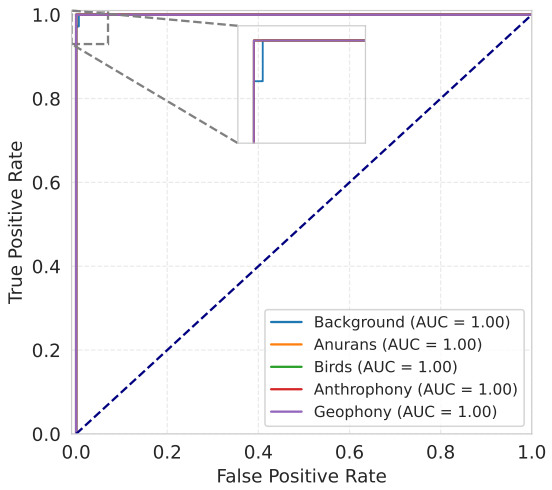
Overall, there is a trade-off between the false-positive rate (FPR) and the true-positive rate (TPR or sensitivity) obtained when we sweep the decision threshold over the OCSVM score s (Figure 5). Each Receiver Operating Characteristic (ROC) curve is obtained by classifying a segment as target whenever $s \geq \tau$ and sweeping the threshold τ across the score range. Curves closer to the upper-left corner indicate better performance, whereas the diagonal corresponds to random guessing (Fawcett, 2006).

The area under the ROC curve (AUC) summarizes performance across all thresholds. The AUC ranges from 0 to 1, with 0.5 indicating chance level and 1.0 a perfect classifier (Fawcett, 2006). Probabilistically, $\text{AUC} = \Pr(s^+ > s^-)$: the probability that a randomly chosen positive (target call) receives a higher score than a randomly chosen negative (non-target). Because the AUC is threshold-independent and invariant to any monotonic transformation of s, it provides a robust, single-number measure of separability for the OCSVM in this setting. It is evident that applying the band-pass filter prior to feature extraction and classification improves AUC and makes the classifier less sensitive to variations in the decision threshold τ (Figure 5b). In other words, the classes become better separated

To build intuition about this separability, we performed dimensionality reduction using Principal Component Analysis (PCA) to project the 1,280-dimensional embedding vectors onto two dimensions. *S. bicolor* call segments are well separated in feature space, facilitating OCSVM classification (Figure 6a). Zooming in on the positive class, it highlights which of the samples were marked as outliers during OCSVM training (Figure 6b). Beyond PCA, other analyses can be performed to evaluate the quality of the embeddings. For instance, compactness (within-group cohesion) of the positive-class vectors are evident when they are clustered using cosine similarity (Figure D.12 in Ap-

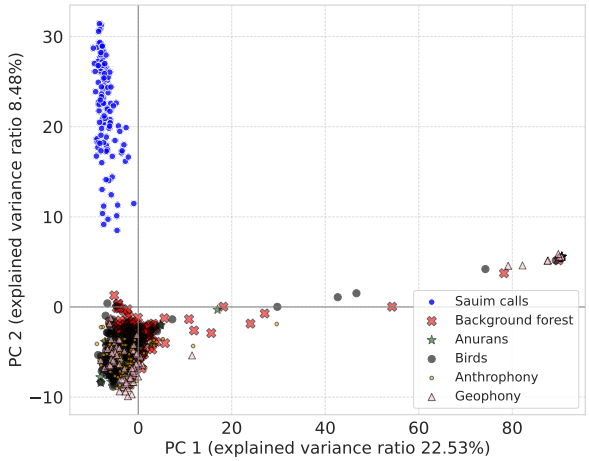


(a) ROC curves before band-pass filtering.

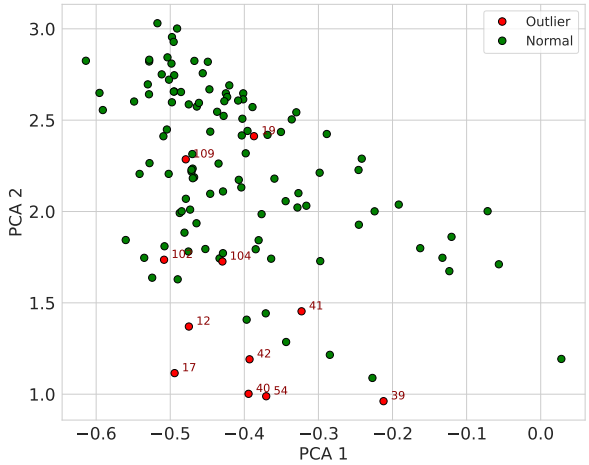


(b) ROC curves after band-pass filtering. The square outlines the area magnified in the inset (upper-left corner).

Figure 5: Comparison of ROC curves for OCSVM detection before and after band-pass filtering; most curves overlap, with filtering yielding a slight up-and-left shift (higher AUC).



(a) All samples, colored by class labels.



(b) Positive class with OCSVM-flagged outliers highlighted (11 outliers).

Figure 6: Two-dimensional projection of call-segment embeddings onto the first two principal components (PCA). Panels: (a) all classes; (b) positive class with outliers.

pendix D). However, a deeper analysis of embedding quality is beyond the scope of this work.

Finally, Table 1 shows the performance of our OCSVM across five metrics—accuracy (Acc), precision (Prec), recall (Rec), weighted F-score (F1), and AUC—evaluated against the five datasets used as negative classes, yielding an open-set scenario. All these metrics were computed on the band-pass–filtered signals (Table 1). The weighted F1 score is computed as

$$F1_{\text{weighted}} = \sum_k \left(\frac{n_k}{\sum_j n_j} \right) \frac{2\, P_k\, R_k}{P_k + R_k}, \qquad (2)$$

where k indexes the classes (here, inlier vs. outlier), n_k is the support (number of true samples) of class k, $P_k = \text{TP}_k/\text{TP}_k{+}\text{FP}_k$ is precision, and $R_k = \text{TP}_k/\text{TP}_k{+}\text{FN}_k$ is recall for class k.

The F1 remains above 96% in all cases, with only minor variation across datasets (Table 1). This stability reflects (i) the strong separability induced by the Filter+Perch+OCSVM pipeline (consistent with the overlapping ROC curves and high AUC), and (ii) the weighting by class support in the F-score, which reduces sensitivity to small fluctuations in the minority class. Overall, the results indicate excellent performance of the proposed pipeline for *S. bicolor* call detection under simulated open-set conditions.

Table 1: Results after applying the band-pass filter. Before filtering, the Background class had precision 0.84 and F1 0.95. The remaining classes were unchanged.

	Acc	Prec	Rec	F1	AUC
Background	0.95	1.00	0.89	0.98	1.00
Anurans	0.97	1.00	0.89	0.97	1.00
Birds	0.98	1.00	0.89	0.98	1.00
anthropophony	0.97	1.00	0.89	0.97	1.00
Geophony	0.96	1.00	0.89	0.96	1.00

Table 2

Band-pass	Acc	Prec	Rec	F1	AUC
yes	0.69	0.86	0.30	0.65	0.83
no	0.65	0.80	0.18	0.57	0.74

6. Use case: a detection example

In our pipeline, 'classification' and 'detection' are tightly coupled: to detect the target species we classify short segments extracted from a long recording. The experiment in this section simulates scanning a long audio record file to detect *S. bicolor* calls. A real deployment requires the classifier to perform well on recordings from new geographic sites not present in training. Although we search for the same species, the forest background and co-occurring fauna can differ substantially across sites, producing a distribution shift in the embeddings/features and thus testing the generalization of our method.

For this detection example, we used a ~10-minute recording from Mindú Park for validation. Note that Mindú Park is geographically distinct from Bosque da Ciência, and the *S. bicolor* groups at the two sites are not in contact. Surroundings also differ, leading to different interference sources in the audio. The recording was human-annotated: the onset of each *S. bicolor* vocalization was time-stamped (Figure 7). We processed the waveform (amplitude normalized to $[-1, 1]$) with a five-second sliding window and a one-second hop (80% overlap), yielding 559 segments (and therefore 559 embedding vectors). The OCSVM then labeled each segment as inlier (sauim) or outlier (non-sauim). In total, 51 detections were registered.

The red trace in Figure 7 marks the detected *S. bicolor* segments. Consecutive windows predicted with the same label are merged into continuous regions (solid red). We also show spectrograms for two contiguous intervals rich in calls, highlighting both the variability in call spectral shapes and changes in background noise—common in long recordings. In this example, we observe a false negative rate (FPR) of 69% and a false positive rate (FNR) of 0.03% (approximately zero).

To quantify detector quality, we matched model predictions to the human annotations using a simple temporal-overlap rule: a call is considered correctly detected if the human-annotated onset time falls within any predicted five-second positive window (equivalently, within ±2.5 s of that window's center). This approximation is sufficient to compare and assess whether band-pass filtering prior to classification improves results.

Figure 8 shows ROC/AUC improvements when filtering is applied prior to feature extraction. These results reinforce that our method can be applied to new recordings from a different geographic region, with a different set of *S. bicolor* individuals and changing background conditions. Finally, the relatively small F1 values are driven by the lower recall (with high precision), indicating a detector that only labels *S. bicolor* when confidence is high—thereby avoiding false positives (Table 2).

In the next section, we extend this cross-site evaluation by analyzing eight recordings from Mindú Park—geographically distant from the site used to train the OCSVM—to assess geographic transferability. Detection counts for these recordings are summarized in Table 3. We then show how to estimate detection probabilities at occupied site by plugging our detector into an occupancy model.

7. Cross-site validation of the OCSVM detector with an occupancy model

Full timestamp annotation of all pied-tamarin vocalizations in PAM recordings is attention-intensive and time-consuming (e.g. the timestamp labels of Section 6). A simpler alternative—counting calls per unit time to estimate call density—still requires listening to entire files. When the objective is merely presence/absence at a site-visit, labeling can be further simplified by early stopping: listen only until the first confirmed vocalization, then stop. Despite its simplicity, these binary labels are sufficient for the occupancy framework presented by MacKenzie et al. (2003).

Our acoustic design follows the single-season occupancy framework of Royle and Link (2006); the key difference is that detections come from the OCSVM applied to acoustic recordings rather than from field observers. Let $y_{it} \in 0, 1$ denote detection at site i on visit t ($t = 1, \ldots, T$), and let $z_i \in 0, 1$ be the latent occupancy state with probability $\Psi = \Pr(z_i = 1)$. If we assume conditional independence of visits given z_i and let

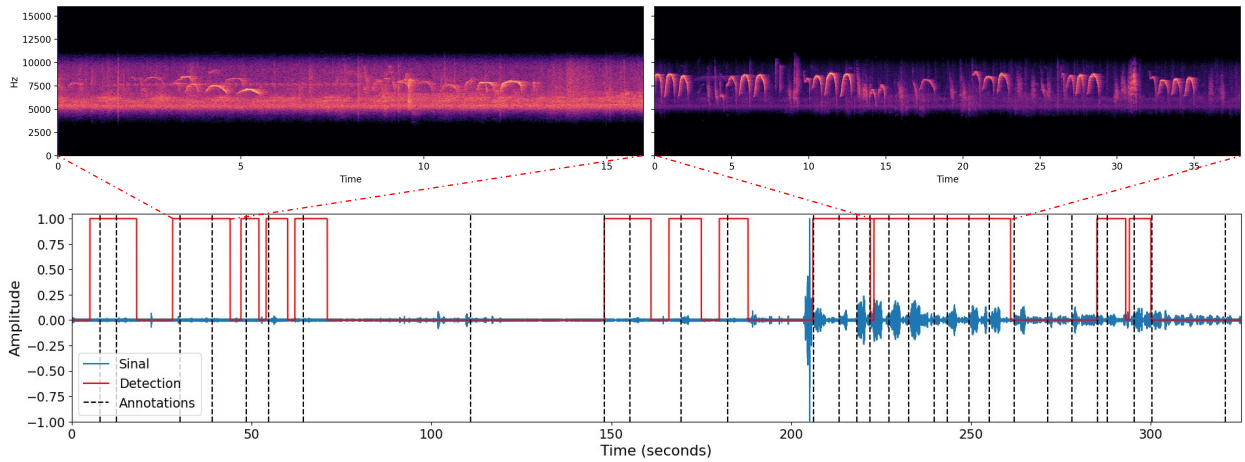


Figure 7: Detection timeline for a representative filtered recording of Mindú park. The waveform is shown with vertical dashed black lines marking the manually annotated onset of each *S. bicolor* call. Red lines denote the sequence of fixed-length time windows classified as inliers by the OCSVM, and spectrograms of the selected windows are displayed above. In this example, the detector yields no false positives but misses some calls (false negatives).

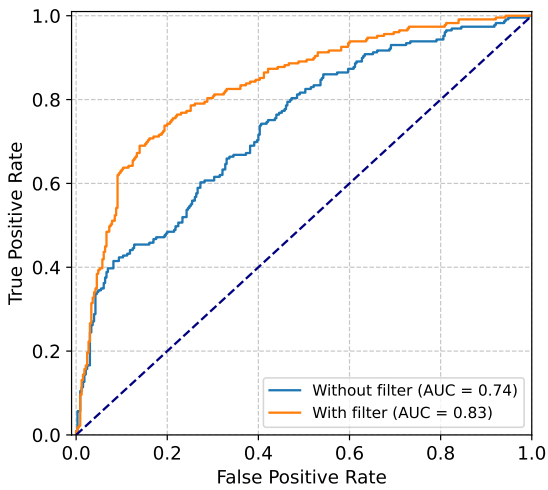


Figure 8: ROC curves for OCSVM-based detection computed on raw (unfiltered) audio and on band-pass–filtered audio. Band-pass filtering consistently improves performance across thresholds, yielding higher true positive rates and a larger area under the curve (AUC). The diagonal line denotes the random-classifier reference.

$y_i = \sum_{t=1}^{T_i} y_{it}$ be the total number of detections at site i across T_i visits, the likelihood is

$$L(\Psi, p_{11}, p_{10} \mid \mathbf{y}) \propto \prod_{i=1}^{R} \Big[\Psi\, p_{11}^{y_i} (1-p_{11})^{T_i - y_i} + (1-\Psi)\, p_{10}^{y_i} (1-p_{10})^{T_i - y_i} \Big], \tag{3}$$

where R is the number of monitored sites and y_i is the total number of positive detections at site i visited T_i times.

Table 3: Detection summary for Mindú Park recordings. Columns correspond to individual recordings $r_1, \ldots, r_8$. W is the number of sliding windows extracted from each recording; D is the number of windows detected as containing an *S. bicolor* call; and y_t indicates whether at least one detection occurred at visit t.

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
W	757	149	109	309	1036	360	419	1320
D	51	0	0	10	261	55	54	112
y_t	1	0	0	1	1	1	1	1

Because Ψ, p_{11}, and p_{10} are unknown, they must be estimated jointly. The probabilities p_{11} and p_{10} depend on the detector's error rates; with automated algorithms these are hard to specify a priori and typically require a hold-out dataset. The Mindú Park dataset contains eight recordings of varying length (Table 3), all labeled as positive (= 1, pied tamarin present), so for these recordings we may take $\Psi \approx 1$ (i.e., treat $z_i \equiv 1$). Under this assumption, the mixture component for unoccupied sites vanishes and the likelihood simplifies to the binomial form

$$L(p_{11} \mid \mathbf{y}) \propto \prod_{i=1}^{R} p_{11}^{y_i} (1-p_{11})^{T_i - y_i}. \tag{4}$$

Note that with only occupied sites we can estimate p_{11}, whereas p_{10} requires data from unoccupied sites.

In this set of experiments, we consider a single site ($R = 1$) visited on eight days ($T = 8$). The recordings come from a different *S. bicolor* group than the one used to train our classifier. Each recording has

varying length and vocalization rate (Table 3). Because the numbers of windows W and detections D_t per visit can vary, we henceforth work with the binary indicator $y_t = \mathbf{1}\{D_t > 0\} \in 0, 1$. Under this reduction, Equation 4 simplifies to $L(p_{11} \mid \mathbf{y}) \propto p_{11}^{y}(1 - p_{11})^{T-y}$, where $y = \sum_{t=1}^{T} y_t$. For computational convenience, we express the corresponding log-likelihood as

$$\ell(p_{11}) = y \log(p_{11}) + (T - y) \log(1 - p_{11}), \qquad (5)$$

which yields the MLE $\hat{p}_{11} = y/T = 0.75$ and the standard error

$$\mathrm{SE}(\hat{p}_{11}) = \sqrt{\frac{\hat{p}_{11}(1 - \hat{p}_{11})}{T}} = 0.15.$$

In our occupancy notation for occupied sites, the miss probability is $p_{01} = Pr(y_t = 0 \mid z = 1) = 1 - p_{11}$. Thus, once we estimate $\hat{p}_{11}$, we immediately obtain $\hat{p}_{01} = 1 - \hat{p}_{11} = 0.25$. Appendix C shows the Maximum-Likelihood estimation (MLE) curve of p_{01}. This helps characterize how well our detector can support a human expert when verifying PAM records.

8. Software package

To make the method accessible for conservation practitioners, we provide an open-source Python package, `sauim-detector`, distributed under the MIT license. The package can be installed via `pip` and runs locally. Given an audio file, the software applies the proposed pipeline—band-pass filtering, Perch embeddings, and OCSVM classification—and generates two outputs: (i) an Audacity-compatible label file (.txt) containing the detection timestamps, and (ii) a new .wav file representing a band-pass–filtered version of the original recording.

The label file can be directly imported into Audacity (`File`→`Import`→`Labels`), enabling analysts to visually inspect spectrograms and confirm *S. bicolor* calls, while the filtered audio provides a cleaner acoustic signal for playback or further analysis (Figure 9). This design facilitates faster manual verification by focusing expert effort on candidate segments rather than entire recordings, thus reducing workload in large PAM datasets.

9. Discussion and conclusions

We presented a practical pipeline for open-set acoustic detection of the pied tamarin (*Saguinus bicolor*), combining a band-pass filter informed by domain knowledge, Perch embeddings, and a One-Class Support Vector Machine (OCSVM). Across multiple non-target datasets that simulate open-set conditions, the approach separated *S. bicolor* calls from other acoustic events and showed that a bird-oriented embedding model (Perch) can transfer effectively to a primate species when paired with a band-pass filter and a one-class classifier. In parallel, we described how the method operates in practice—scanning long recordings with sliding windows to enable either streaming detection or offline batch processing over large archives—so that detections become time-localized and directly usable in monitoring workflows.

In our cross-site detection evaluation we obtained FPR = 0.03, FNR = 0.69, TPR = 0.31, and TNR = 0.97. These results indicate high specificity but limited sensitivity. This suggests that Perch+OCSVM transfers across a different *S. bicolor* populations—despite possible variation in vocalization patterns—while remaining conservative. These figures should be interpreted as approximate because onset timestamps may be slightly misaligned with the detector's analysis windows, which can affect event matching metrics. In addition, acoustic signals attenuate with distance from the recorder—particularly at higher frequencies—reducing the signal-to-noise ratio and the fidelity of distant calls; this tends to increase FNR and can degrade PAM performance.

In large PAM datasets spanning many hours of audio, it is common to pre-screen recordings with an automated detector to generate candidate timestamps. A human expert then reviews only these candidates to confirm that each positive detection corresponds to the target species (e.g., pied tamarin), thereby reducing manual time and effort. For this verification workflow, a very desirable property is a low false-negative rate (high sensitivity) to avoid missing all calls and, consequently, labeling an entire recording as "no species present". In this context, the software package presented in Section 8 also has the potential to substantially optimize both human effort and financial resources in field surveys aimed at detecting the species in new areas, given that detecting the species often requires several days of intensive fieldwork and long, exhausting walks, even within small forest fragments.

Building on these detections, we link the detector outputs to a simplified single-site occupancy framework. Assuming Mindú Park is occupied ($\Psi = 1$), the MLE has the closed form $\hat{p}_{11} = y/T$, where y is the number of visits with at least one detection across T visits; the corresponding false-negative probability is $p_{01} = 1 - p_{11} = 0.25$, indicating good sensitivity of

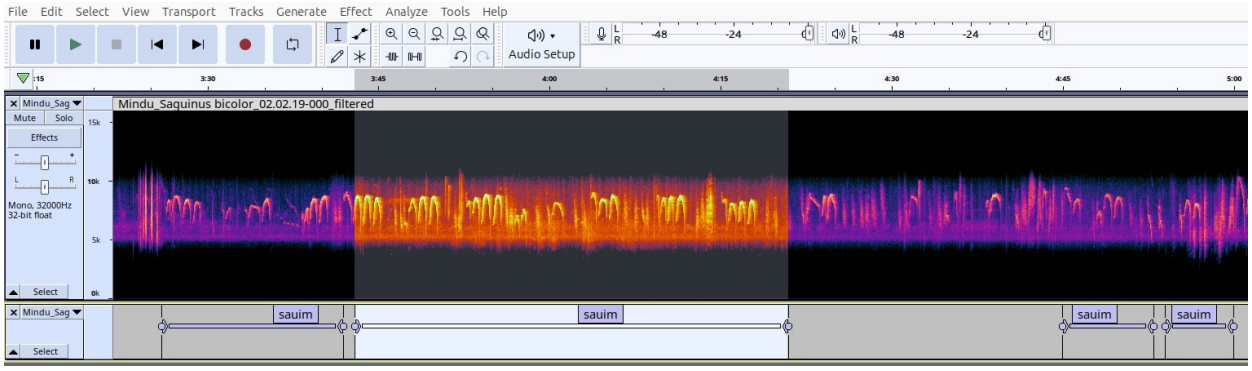


Figure 9: Example Audacity view after running `sauim-detector` and import the lables file. The spectrogram of the band-pass–filtered audio (top) and the imported label track highlighting candidate *S. bicolor* detections (bottom), enabling faster manual verification.

the detection pipeline. Although simple, this example shows how the complete pipeline—from band-pass filtering and Perch embeddings to OCSVM decisions—can be deployed end-to-end for occupancy inference.

Under manual inspection, false positives are comparatively easy to verify—brief spot-checks suffice, avoiding a full review of the dataset. False negatives are more consequential: an analyst may prematurely infer absence and stop screening, potentially biasing ecological inference. Because calling recurs over time, the cumulative probability of detection increases across visits, partly offsetting this risk. Nevertheless, an important caveat applies: the Mindú Park recordings were collected manually and deliberately close to the acoustic source (near *S. bicolor*); in PAM programs with fixed-position recorders, the source is not always close to the microphone, which can reduce SNR and limit the transferability of these detection-probability estimates.

There are, however, clear avenues for improvement. Although Perch transferred well to pied tamarin, exploring alternative or complementary embedding models, along with modest fine-tuning on this primate calls, may further improve detection performance. The influence of closely related species—such as the golden-handed tamarin (*Saguinus midas*)—should be explicitly evaluated. Incorporating active learning (Dumoulin et al., 2025) or semi-supervised learning could also improve robustness and reduce annotation effort while adjusting the decision boundary to novel acoustic environments. Finally, scaling to multi-site or multi-season designs with covariates such as seasonal variations and habitat variables could help to improve detections.

Acknowledgements

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES-PROEX), Financing Code 001, and by the Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM, through the PDPG/CAPES and POSGRAD 2025/2026 projects. We also thank FAPEAM for its support through the project titled "Diferentes Abordagens Computacionais para Monitoramento Ecoacústico Autônomo da Região Amazônica", linked to Public Call No. 013/2022—Produtividade em CT&I. JGC thanks OpenAI for its support through the partnership established with its representative, Nicolas Robinson Andrade, as part of the AmazonGPT project. TVS thanks The Rufford Foundation (24762-1 and 37915-2), The National Geographic Society (EC-419R-18) and the International Primatological Society for funding fieldwork.

References

Albernaz, A.L., Thalês, M.C., Gordo, M., Lagroteria, D., Sobroza, T.V., Magnusson, W.E., Fearnside, P.M., Jerusalinsky, L., de Azevedo, R.B., Castro, R.B., Campista, D., Spironello, W.R., Noronha, M., 2026. Conservation of an endangered amazonian primate: priority areas for the pied tamarin (saguinus bicolor) in manaus, brazil. Journal for Nature Conservation 89, 127069. doi:10.1016/j.jnc.2025.127069.

Ayres, J.M., Mittermeier, R.A., Constable, I.D., 1982. Brazilian tamarins on the way to extinction? Oryx 16, 329–333. doi:10.1017/S0030605300017786.

Colonna, J., Peet, T., Ferreira, C.A., Jorge, A.M., Gomes, E.F., Gama, J., 2016. Automatic classification of anuran sounds using convolutional neural

networks, in: Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, Association for Computing Machinery. p. 73–78. doi:10.1145/2948992.2949016.

Crunchant, A.S., Borchers, D., Kühl, H., Piel, A., 2020. Listening and watching: Do camera traps or acoustic sensors more efficiently detect wild chimpanzees in an open habitat? Methods in Ecology and Evolution 11, 542–552.

Dumoulin, V., Stretcu, O., Hamer, J., Harrell, L., Laber, R., Larochelle, H., van Merriënboer, B., Navine, A., Hart, P., Williams, B., Lamont, T.A.C., Razak, T.B., Team, M.C.R., Brodie, S., Doohan, B., Eichinski, P., Roe, P., Schwarzkopf, L., Denton, T., 2025. The search for squawk: Agile modeling in bioacoustics. doi:10.48550/arXiv.2505.03071, arXiv:2505.03071.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874. doi:10.1016/j.patrec.2005.10.010.

Fernandes, L.S., Sobroza, T.V., Rodrigues, E.C., Gordo, M., 2024. Germination success of seeds ingested by saguinus bicolor. Studies on Neotropical Fauna and Environment 60, 179–190.

Geng, C., Huang, S.J., Chen, S., 2021. Recent advances in open set recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 3614–3631. doi:10.1109/TPAMI.2020.2981604.

Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Scientific Reports 13, 22876. doi:10.1038/s41598-023-49989-z.

Google Research, 2025. Perch v8 bioacoustic model. Kaggle Models, Bird Vocalization Classifier (v8). Https://www.kaggle.com/models/google/bird-vocalization-classifier/8, Accessed on 13 August 2025.

Gordo, M., Calleia, F.O., Vasconcelos, S.A., Leite, J.J.F., Ferrari, S.F., 2013. The Challenges of Survival in a Concrete Jungle: Conservation of the Pied Tamarin (Saguinus bicolor) in the Urban Landscape of Manaus, Brazil. Springer New York. pp. 357–370. doi:10.1007/978-1-4614-8839-2_23.

Gordo, M., Röhe, F., Vidal, M.D., Subirá, R., Boubli, J.P., Mittermeier, R.A., Jerusalinsky, L., 2021. *Saguinus bicolor* (amended version of 2019 assessment). doi:10.2305/IUCN.UK.2021-1.RLTS.T40644A192551696.en.

ICMBio, 2011. Plano de Ação Nacional para conservação do sauim-de-coleira (PAN sauim-de-coleira). https://www.gov.br/icmbio/pt-br/assuntos/biodiversidade/pan/pan-sauim-de-coleira. Accessed on 7 September 2025.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. Birdnet: A deep learning solution for avian diversity monitoring. Ecological Informatics 61, 101236. doi:10.1016/j.ecoinf.2021.101236.

Khan, S.S., Madden, M.G., 2014. One-class classification: taxonomy of study and review of techniques. The Knowledge Engineering Review 29, 345–374. doi:10.1017/S026988891300043X.

Lagroteria, D., Cavalcante, T., Zuquim, G., Röhe, F., Medeiros, A.S., Hrbek, T., Gordo, M., 2024. Assessing the invasive potential of saguinus midas in the extent of occurrence of the critically endangered saguinus bicolor. Frontiers in Conservation Science 5, 1426488.

Laidre, K.L., Zahn, M.J., Simon, M., Ladegaard, M., Stafford, K.M., Phillips, E., Moon, T., Stern, H.L., Cohen, B., 2024. Narwhal (Monodon monoceros) associations with greenland summer meltwater release. Ecosphere 15, e70024. doi:10.1002/ecs2.70024.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., Franklin, A.B., 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84, 2200–2207. doi:10.1890/02-3090.

Marques, H.O., Swersky, L., Sander, J., Campello, R.J.G.B., Zimek, A., 2023. On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles. Data Mining and Knowledge Discovery 37, 3614–3631. doi:10.1007/s10618-023-00931-x.

Miksis-Olds, J.L., Tyack, P.L., 2009. Manatee (Trichechus manatus) vocalization usage in relation to environmental noise levels. The Journal of the Acoustical Society of America 125, 1806–1815. doi:10.1121/1.3068455.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,

P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Piczak, K.J., 2015. ESC: Dataset for Environmental Sound Classification, in: Proceedings of the 23rd Annual ACM Conference on Multimedia, pp. 1015–1018. doi:10.1145/2733373.2806390.

Royle, J.A., Link, W.A., 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87, 835–841. doi:10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2.

Scheirer, W.J., Rocha, A., Sapkota, A., Boult, T.E., 2013. Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 1757–1772. doi:10.1109/TPAMI.2012.256.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Computation 13, 1443–1471. doi:10.1162/089976601750264965.

Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J., 1999. Support vector method for novelty detection, in: Solla, S., Leen, T., Müller, K. (Eds.), Advances in Neural Information Processing Systems, MIT Press.

Sobroza, T.V., Bueno, A.S., Prestes, A., Masseli, G.S., Appel, G., Martins, G.M., Lima, H., Kaefer, I.L., dos Santos, J.V.C., Ramos, J., et al., 2025. Protocolo de amotragem da biodiversidade via monitoramento acústico passivo. Educamazônia - Educação, Sociedade e Meio Ambiente 1, 441–472.

Sobroza, T.V., Cerqueda, L.S., Simões, P.I., Gordo, M., 2017. Vocal repertoire and its behavioral contexts in the pied tamarin, saguinus bicolor. International Journal of Primatology 38, 642–655. doi:10.1007/s10764-017-9971-z.

Sobroza, T.V., Gordo, M., Dunn, J.C., Pequeno, P.A.C.L., Naissinger, B.M., Barnett, A.P.A., 2024. Pied tamarins change their vocal behavior in response to noise levels in the largest city in the amazon. American Journal of Primatology 86, e23606. doi:10.1002/ajp.23606.

Sobroza, T.V., Gordo, M., Pequeno, P.A.C.L., Dunn, J.C., Spironello, W.R., Rabelo, R.M., Barnett, A.P.A., 2021. Convergent character displacement in sympatric tamarin calls (Saguinus spp.). Behavioral Ecology and Sociobiology 75, 88. doi:10.1007/s00265-021-03028-x.

Vidal, M.D., Cintra, R., 2006. Effects of forest structure components on the occurence, group size and density of groups of bare-face tamarin (saguinus bicolor - primates: Callitrichinae) in central amazonia. Acta Amazonica 36, 237–248. doi:10.1590/S0044-59672006000200014.

Williams, B., van Merriënboer, B., Dumoulin, V., Hamer, J., Fleishman, A.B., McKown, M., Munger, J., Rice, A.N., Lillis, A., White, C., Hobbs, C., Razak, T., Curnick, D., Jones, K.E., Denton, T., 2025. Using tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. Philosophical Transactions of the Royal Society B: Biological Sciences 380, 20240280. doi:10.1098/rstb.2024.0280.

Wood, C.M., Barceinas Cruz, A., Kahl, S., 2023. Pairing a user-friendly machine-learning animal sound detector with passive acoustic surveys for occupancy modeling of an endangered primate. American Journal of Primatology 85, e23507.

Xavier, R.d.S., Gosset, M., Maciel, T.F., Bicudo, T., Nascimento, L.d., Ramalho, E., Fleischmann, A., 2024. Measuring amazon rainfall intensity with sound recorders. Geophysical Research Letters 51, e2024GL108210.

Appendix A. Species

Bird dataset species composition: *Cacicus haemorrhous, Cercomacroides nigrescens, Chiroxiphia pareola, Chrysomus icterocephalus, Glaucidium hardyi, Graydidascalus brachyurus, Hemitriccus minor pallens, Herpsilochmus dorsimaculatus, Hypocnemis flavescens, Ibycter americanus, Leistes militaris, Myiopagis gaimardii, Myrmoborus lugubris stictopterus, Neopelma chrysocephalum, Pharomachrus pavoninus, Pionus fuscus, Pseudopipra pipra, Ramphocelus carbo, Rupicola rupicola, Schiffornis olivacea, Sublegatus obscurior, Thamnophilus amazonicus cinereiceps, Thamnophilus murinus, Todirostrum pictum, Tyrannus melancholicus, Zimmerius acer*

Anuran dataset species composition: *Adenomera hylaedactyla, Hyla minuta, Adenomera andreae, Ameerega trivittata, Osteocephalus oophagus, Rhinella granulosa, Scinax ruber, Hypsiboas cinerascens,*

Brachycephalus ephippium, Aplastodiscus albofrenatus, Aplastodiscus albosignatus, Aplastodiscus perviridis, Dendropsophus anceps, Dendropsophus elegans

Appendix B. Confusion matrices

(a) Anurans

(b) Birds

(c) anthropophony

(d) Geophony and insects

Figure B.10: Confusion matrices for the OCSVM with band-pass filtering. Each matrix reports TP, FP, TN, and FN counts on the evaluation split at the default decision threshold.

Appendix C. Maximum likelihood estimation

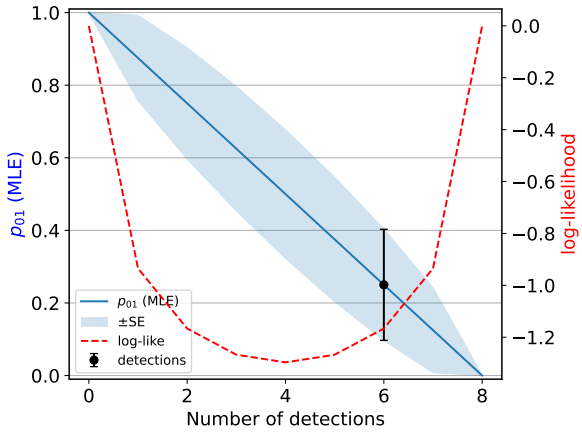


Figure C.11: MLE of $\hat{p}_{01} = 1 - \hat{p}_{11}$ as a function of the number of detections k. The blue curve shows $\hat{p}_{01}(k)$ with a shaded ±1 standard-error band. The black point highlights the estimate at the observed detections $k = \sum_i \mathbf{1}\{y_i > 0\}$, with its error bar. The red dashed line (right axis) shows the maximized log-likelihood $\max_{p_{11}} \log L(k \mid p_{11})$ as a function of k.

Appendix D. Cluster of embeddings vectors

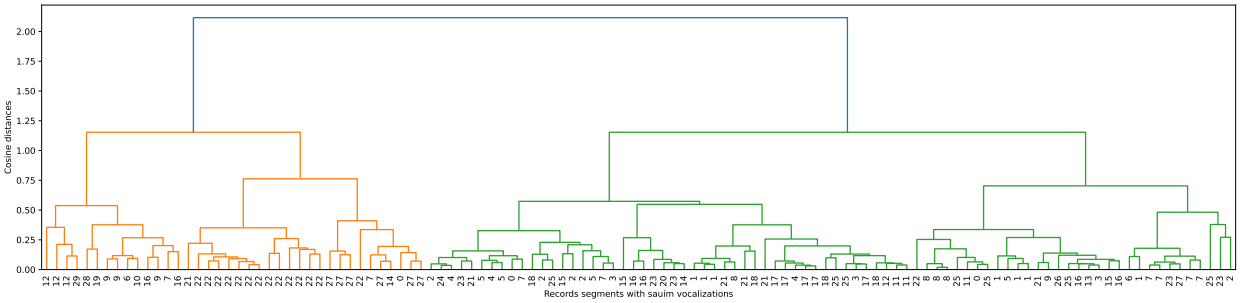


Figure D.12: Hierarchical clustering dendrogram of positive-class *S. bicolor* embeddings (cosine distance, average linkage). Each leaf is labeled with the integer ID of the recording from which the call segment was extracted. Segments from the same recording tend to cluster together, likely reflecting the same individual and shared background noise.