Evaluating Vector Fusion Techniques for Unsupervised Soundscape Retrieval using Deep Embeddings

Andrés D. Peralta[1], **Alejandro C. Frery**[2], **Jie Xie**[3], **Juan G. Colonna**[1,2]

[1] Instituto de Computação - Universidade Federal do Amazonas (UFAM)
[2] School of Mathematics and Statistics - Victoria University of Wellington (VUW)
[3]School of Artificial Intelligence - Nanjing Normal University

{andres, juancolonna}@icomp.ufam.edu.br
xiej8734@gmail.com
alejandro.frery@vuw.ac.nz

Abstract. *The efficient storage, management, and retrieval of vast volumes of bioacoustic data represent a critical bottleneck for long-term biodiversity monitoring. To address this challenge, this work proposes an unsupervised soundscape retrieval system that integrates high-dimensional embeddings from the pre-trained Perch V2 model with a vector fusion strategy that handles variable-length recordings. We systematically evaluate retrieval efficiency and accuracy by comparing two retrieval algorithms and four vector fusion techniques on a vector database. The methodology was validated using a multitaxonomic dataset comprising birds, mammals, and amphibians. A case study involving eleven species of conservation interest shows that the system significantly outperforms traditional MFCC-based approaches, offering a scalable, robust solution for autonomous biodiversity inventorying.*

Keywords: *Biodiversity Monitoring; Vector Fusion; Vector Databases; Acoustic Similarity Retrieval;*

1. Introduction

Soundscape analysis is fundamental for bioacoustic and ecoacoustic monitoring because it exploits low-cost passive sensors and recent advances in machine learning [Pijanowski et al., 2011, Bianco et al., 2019]. However, the field faces significant challenges due to large volumes of unlabeled data, overlapping acoustic sources, and environmental noise variability [Ravanelli et al., 2014]. Traditional feature-based retrieval methods, such as Mel-frequency cepstral coefficients (MFCCs), often fail to capture the complexity of acoustic scenes [Wichern et al., 2010].

To address these limitations, we propose an efficient soundscape retrieval system based on acoustic similarity and a vector database. The system integrates high-dimensional embeddings extracted from Google's Perch V2 [van Merriënboer et al., 2025], a pre-trained bioacoustic model based on the EfficientNet-B1 neural network architecture. The temporal sequences of these embeddings are aggregated using a fusion strategy to condense temporal information into a single representative vector suitable for storage in a vector database.

Our primary contribution lies in the comparison of four vector fusion strategies integrated within a vector database, specifically the `VectorDB` framework. We also conducted a comparative analysis between the Hierarchical Navigable Small World (HNSW)

and the In-Memory ExactNN Index (IMENN) algorithms to identify the most efficient approach for similarity-based audio retrieval. Furthermore, the system's practical applicability is validated through a real use case involving eleven species across different conservation statuses, highlighting its potential for biodiversity monitoring and conservation decision-making.

2. Problem Description

To extract temporal features from raw audio recordings, we model each signal as a one-dimensional time series $s(t) \in \boldsymbol{R}^T$, where T is the total number of samples and $t \in \{1, 2, \ldots, T\}$. The signal is sampled at a constant rate f_s (in Hz) and segmented into non-overlapping windows of fixed duration $\Delta = 5$ s. Each window contains $N = f_s \cdot \Delta$ samples, and the total number of full windows that can be extracted is $M = \lfloor T/N \rfloor$.

Let $s_i \in \boldsymbol{R}^N$ denote the i-th segment of the signal, with $i = 1, \ldots, M$. Each segment is processed by a feature extraction function $f\colon \boldsymbol{R}^N \to \boldsymbol{R}^d$, which produces a fixed-size real-valued feature vector $x_i \in \boldsymbol{R}^d$, where $d = 1280$ for the Perch V2 model and N is typically a 5-second segment. The full audio recording is thus represented by a feature matrix $\mathbf{X} \in \boldsymbol{R}^{d \times M}$, constructed by stacking the feature vectors column-wise $\mathbf{X} = [x_1\ x_2\ \cdots\ x_M]$.

Since the input audio duration is variable, the number of columns M in matrix $\boldsymbol{X}$ also varies across recordings. Thus, we face the challenge of obtaining a fixed-length representation suitable for indexing and comparison. Hence, we apply a fusion technique over the time dimension (columns) of $\mathbf{X}$ and obtain a global embedding $\overline{x} \in \boldsymbol{R}^d$, which condenses the full audio recording into a single representative vector. This step enables consistent downstream processing regardless of the original signal length.

3. Retrieval Algorithms and Evaluation Metrics

The vector database VectorDB is built upon two search algorithms. The HNSW algorithm is used for Approximate Nearest Neighbors (ANN) search based on graph structures [Malkov and Yashunin, 2020]. This algorithm extends the classical nearest-neighbor method by introducing hierarchical graph layers to enable efficient search. Additionally, HNSW calculates similarity using the Euclidean distance between embedding vectors.

In contrast, the *In-Memory ExactNN Index* (IMENN) is an exact k-NN algorithm. During the indexing phase, the set of embedding vectors $X = \{x_1, x_2, \ldots, x_n\}$ is stored as a flat table in memory. To perform a query, a new embedding vector (x_j) is used as input, and similarity is computed between the query and all stored vectors using the Euclidean distance. This distance metric was selected because it is a standard, computationally efficient baseline natively optimized by both HNSW and IMENN implementations within the VectorDB framework. The vectors with the highest similarity scores are then returned as the result set.

The performance of an audio retrieval system implemented over a vector database can be evaluated using a pair of metrics that reflect both the quality and efficiency of the returned results. Commonly used metrics in the literature include *HitRate-k* ($H@k$) and query response time measured in milliseconds. The $H@k$ metric measures the proportion of truly relevant results among the top k results returned by the system, and is defined

as [Krauss et al., 2023]:

$$H@k = \frac{1}{k} \sum_{i=1}^{k} r(i), \tag{1}$$

where k denotes the number of retrieved results, and $r(i)$ is an indicator function that returns 1 if the i-th result is relevant and 0 otherwise. Low $H@k$ values indicate a higher proportion of irrelevant results. In this study, we evaluate both $H@1$ and $H@5$.

Additionally, we use the average query time and its standard deviation as proxies for system efficiency, helping assess its feasibility in interactive search environments, such as web-based services with graphical user interfaces (GUIs), where low latency is critical.

4. Related Work

The increasing impact of machine learning in bioacoustics has driven research into the classification and detection of acoustic events [Bjorck et al., 2019]. Previous studies have explored various retrieval systems, such as CNN-GRU models with Mel-spectral descriptors [Sert and Başbuğ, 2019] and semi-supervised *Deep Hashing* architectures using pre-trained models like VGGish [Jati and Emmanouilidou, 2020]. Additionally, Ghani et al. [2023] showed that embeddings extracted from pre-trained bioacoustic classifiers outperform general audio models in generalizing to novel classes, highlighting the potential of transfer learning for ecoacoustic tasks.

Despite these advances, existing methods face significant challenges. Approaches such as deep supervised hashing [Chen et al., 2019] and product quantization [Liang et al., 2023] are often computationally expensive, limiting their use in large-scale or real-time monitoring scenarios. Furthermore, while vector-based retrieval is well established in domains such as music [Wang, 2003], its systematic application in ecoacoustics is still in its early stages.

Recent research has addressed these gaps through feature fusion techniques to improve species classification and the selection of salient representations to enhance species recognition [Xie and Zhu, 2023, Tolkova, 2019], but lacks integration with vector databases. Our work extends these feature fusion concepts by integrating them into a vector database, enabling efficient similarity retrieval across massive unlabeled datasets.

5. Materials and Methods

5.1. Datasets

This study utilized the `BirdCLEF 2025` dataset, a multitaxonomic bioacoustic collection compiled from diverse geographic locations [Klinck et al., 2025]. The corpus comprises $28\,564$ audio recordings in `OGG` format with variable sampling rates, spanning 206 species of birds, amphibians, and mammals. The recordings exhibit significant temporal diversity, with durations ranging from $0.5\,\mathrm{s}$ to approximately $1774\,\mathrm{s}$ (median = $22.6\,\mathrm{s}$). Although some recordings may contain vocalizations from multiple species (up to two), categorized as primary and secondary labels, the secondary labels are scarce and frequently incomplete. For this reason, only the primary labels were considered in this study.

5.2. Proposed Method

Given that bioacoustic recordings often exhibit variable durations and heterogeneous acoustic content, we hypothesize that fusing embedding feature vectors extracted from sequences of acoustic frames captures the overall acoustic signature of a recording more effectively than extracting a single embedding from the entire audio signal.

While previous studies have explored aggregation methods for traditional acoustic features [Xie and Zhu, 2023], our proposal compares four distinct vector fusion strategies: average, weighted average, sum, and max pooling. These operators aim to condense the sequence of segment-level embedding vectors into a single representative vector per recording, regardless of its duration, thereby facilitating similarity-based retrieval. Under this approach, we expect high similarity between recording-level vectors that share acoustic characteristics, whether they belong to the same species or originate from the same environment or recording session.

Our method consists of four main stages, which are visually grouped into three colored blocks in Figure 1. First, corresponding to the gray Extraction block, data preprocessing is performed where all audio recordings were resampled to $32\,\mathrm{kHz}$, and the stationary Noise Reduce (NR) algorithm[1] was applied. The NR estimates a noise profile from the first $5\,\mathrm{s}$ of each recording via STFT ($N_{\mathrm{FFT}} = 1024$, $50\,\%$ overlap), and builds a frequency-dependent binary mask from the per-band mean and standard deviation, applied to the full spectrogram before inverse STFT. Each recording was then divided into 5-second non-overlapping windows; shorter recordings were excluded from the analysis to maintain a consistent temporal window.

Following this preprocessing phase, in the blue Extraction block, these segments and the query audio are processed via the pretrained Perch V2 model to generate 1280-dimensional embedding vectors. Next, in the green Fusion block, the set of embedding vectors for each recording was aggregated into a single representative vector using one of four fusion strategies. Finally, representing the purple Retrieval block, these fused vectors and their metadata were indexed in a VectorDB database[2], enabling similarity searches through the HNSW and IMENN algorithms. The HNSW index was configured with $M = 32$, $\texttt{efConstruction} = 200$, and $\texttt{efSearch} = 50$.

As a baseline, we implemented traditional MFCCs with 128 coefficients per segment. These features were extracted using a 1024-point FFT with a $50\,\%$ hop length and Hann windowing. The audio was divided into 5-second segments with a 1-second hop length to ensure a fair comparison with our proposed method. Further details on vector merging and evaluation configurations are provided in Sections 5.3 and 5.4.

5.3. Feature Fusion Techniques

To represent each recording of M segments as a fixed-size vector, we evaluated four pooling aggregation techniques applied to the feature matrix $\mathbf{X} \in \boldsymbol{R}^{d \times M}$, where $d = 1280$ for Perch V2 and $d = 128$ for MFCCs:

- **Average Pooling:** Computes the arithmetic mean over the segments as $\bar{x}_{\text{avg}} = M^{-1} \sum_{i=1}^{M} \mathbf{X}_{:,i} \in \boldsymbol{R}^{d}$.

[1] `https://zenodo.org/records/3243139`
[2] `https://github.com/jina-ai/vectordb`

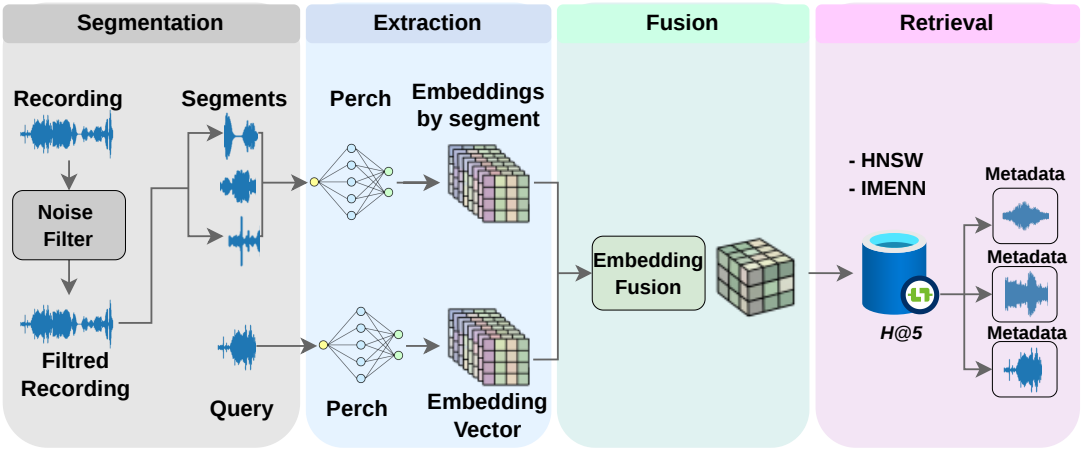

Figure 1. Proposed method for bioacoustic similarity retrieval of soundscapes.

- **Weighted Average Pooling:** Prioritizes acoustically prominent events by weighting each segment x_i by its Root Mean Square (RMS) energy w_i: $\bar{x}_{\text{wavg}} = \left(\sum_{i=1}^{M} w_i\right)^{-1} \sum_{i=1}^{M} w_i \cdot \mathbf{X}_{:,i} \in \boldsymbol{R}^d$.
- **Sum Pooling:** Aggregates total acoustic information across time as $\bar{x}_{\text{sum}} = \sum_{i=1}^{M} \mathbf{X}_{:,i} \in \boldsymbol{R}^d$.
- **Max Pooling:** Highlights the most dominant features by selecting the maximum value along each dimension as $\bar{x}_{\max} = \max_{i=1,\ldots,M} \mathbf{X}_{:,i} \in \boldsymbol{R}^d$.

The resulting one-dimensional vectors ensure computational efficiency and compatibility within the VectorDB. These fusion strategies were applied identically to both deep embeddings and the MFCC baseline to ensure a fair comparison.

5.4. Evaluation Methodologies

The evaluation was structured into two experimental stages. Stage 1 compared different retrieval techniques to identify the most effective fusion strategy. For this purpose, a recording-level split was applied per species, where $70\,\%$ of the recordings were stored in the database and the rest was used as queries. This setup ensures that no temporal or source overlap exists between the sets, preventing information leakage and assessing the system's ability to retrieve acoustically similar segments from distinct recordings. Feature fusion was applied consistently to both the database entries and the query vectors. The dataset for this stage comprised $130\,245$ segments in the database and $57\,594$ query segments.

Stage 2 presented a use case designed to evaluate the system's functional utility. In this scenario, a user uploads an unlabeled recording, and the system returns acoustically similar results along with associated metadata, such as species, location, and source. This use case focuses on eleven representative species from the `BirdCLEF+ 2025` dataset. These species, which include *Colostethus inguinalis*, *Lithobates vaillanti*, *Orophus conspersus*, *Andinobates opisthomelas*, *Alouatta seniculus*, *Nyctibius griseus*, *Setophaga pitiayumi*, *Falco sparverius*, *Sicalis flaveola*, *Tapera naevia* and *Megascops choliba*, were selected based on their conservation status according to The IUCN Red List of Threatened Species [2025]. Finally, geographic metadata (latitude and longitude) was processed using the `geopy` library and the Nominatim service to generate spatial distributions.

Table 1. Species-level retrieval comparison between Perch V2 embeddings and MFCCs across four vector fusion strategies. Δ_{wavg} reports the difference, in percentage points, of $H@1_{\mathbf{s}}$ relative to the Weighted Average. Significance from the paired exact McNemar test with Holm–Bonferroni correction over all pairwise comparisons within each algorithm ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). $t \pm \sigma$ denotes the average query time and its standard deviation in ms.

		HNSW				IMENN			
Vectors	**Technique**	$H@1_{\text{s}}$	$H@5_{\text{s}}$	Δ_{wavg} (pp)	$t \pm \sigma$	$H@1_{\text{s}}$	$H@5_{\text{s}}$	Δ_{wavg} (pp)	$t \pm \sigma$
Perch V2	Avg Pool	0.78	0.89	-6.0^{***}	19 ± 0.17	0.61	0.71	-5.0^{***}	$\mathbf{58 \pm 6.86}$
	Weighted Avg	**0.84**	**0.93**	—	21 ± 0.22	**0.66**	**0.75**	—	58 ± 6.90
	Sum	0.76	0.88	-8.0^{***}	$\mathbf{16 \pm 0.20}$	0.47	0.56	-19.0^{***}	59 ± 6.79
	Max Pool	0.79	0.90	-5.0^{***}	17 ± 0.19	0.62	0.72	-4.0^{***}	59 ± 6.79
MFCCs	Avg Pool	0.24	0.40	-3.0^{**}	5 ± 0.05	0.20	0.35	-3.0^{**}	8 ± 3.37
	Weighted Avg	0.27	0.45	—	6 ± 0.06	0.23	0.39	—	9 ± 3.41
	Sum	0.24	0.40	-3.0^{**}	4 ± 0.05	0.15	0.27	-8.0^{***}	7 ± 3.62
	Max Pool	0.21	0.36	-6.0^{***}	5 ± 0.05	0.18	0.33	-5.0^{***}	8 ± 4.22

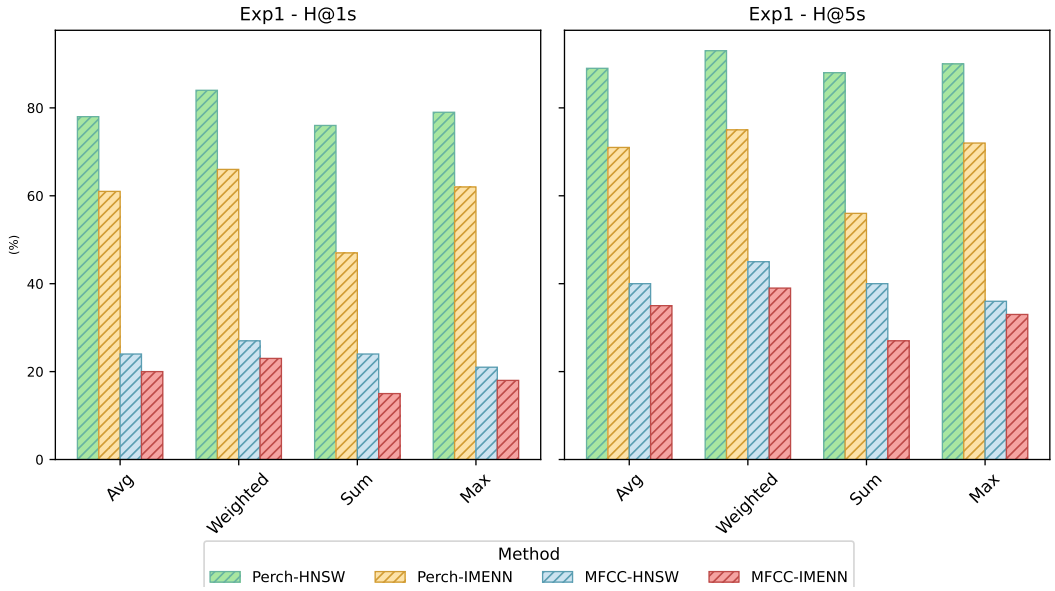

Figure 2. Performance comparison of vector fusion techniques and retrieval algorithms using the $H@k_{\mathbf{s}}$ metrics. Subfigures (a) and (b) represent $H@1_{\mathbf{s}}$ and $H@5_{\mathbf{s}}$ accuracy scores, respectively, expressed as percentages.

6. Results

The results in Table 1 and Figure 2 confirm that the embeddings extracted using the Perch V2 model significantly outperform traditional MFCCs across all configurations. The combination of Weighted Average Pooling and the HNSW algorithm achieved the highest performance, reaching $H@1_s = 0.84$ and $H@5_s = 0.93$. The paired exact McNemar test with Holm–Bonferroni correction confirms that the Weighted Average is statistically superior to all alternative fusion strategies for Perch V2, with a $p < 0.001$ across all pairwise comparisons, and effect sizes between 5 and 8 percentage points in $H@1_s$. The same hierarchy holds for the MFCC baseline, although with smaller margins and a $p < 0.01$,

Table 2. Species information and geographic distribution for the new bioacoustic records. The table summarizes the taxonomic class, data source, IUCN Red List category, and number of recordings retrieved for each newly identified species.

Species	# Recordings	Taxonomic Class	Source	Red List Category	Location
Colostethus inguinalis	2	Amphibia	XC (2)	Least Concern	Col
Lithobates vaillanti	2	Amphibia	XC (2)	Least Concern	Ni
Orophus conspersus	4	Insecta	XC (4)	Least Concern	Col
Andinobates opisthomelas	10	Amphibia	iNat (10)	Vulnerable	Col
Alouatta seniculus	23	Mammalia	XC (1) iNat (22)	Least Concern	Br, Col
Nyctibius griseus	124	Aves	XC (32) iNat (92)	Least Concern	Ar, Bo, Br, Col, Ec, Py, Pe, Ve
Setophaga pitiayumi	298	Aves	XC (126) iNat (172)	Least Concern	Ar, Bo, Br, Col, CR, Ec, Gt, Hn, Mx, Ni, Pa, Py, Pe, Uy, Us, Ve
Falco sparverius	312	Aves	XC (26) iNat (283) CSA (3)	Least Concern	Ar, Bo, Br, Ca, Ch, Col, CR, Ec, Gt, Mx, Ni, Pe, Py, Us, Ve
Sicalis flaveola	323	Aves	XC (202) iNat (85) CSA (36)	Least Concern	Ar, Bo, Br, Col, Ec, Ha, Py, Pe, Uy, Ve
Tapera naevia	379	Aves	XC (176) iNat (203)	Least Concern	Ar, Bo, Br, Col, CR, Ec, Gt, Hn, Mx, Ni, Pa, Py, Pe, Uy, Ve
Megascops choliba	401	Aves	XC (116) iNat (285)	Least Concern	Ar, Bo, Br, Col, CR, Ec, Gt, Hn, Ni, Pa, Py, Pe, Uy, Ve

Location codes: Ar = Argentina, Bo = Bolivia, Br = Brazil, Ca = Canada, Ch = Chile, Col = Colombia, CR = Costa Rica, Ec = Ecuador, Gt = Guatemala, Ha = Havaí, Hn = Honduras, Mx = Mexico, Ni = Nicaragua, Pa = Panama, Py = Paraguay, Pe = Peru, Uy = Uruguay, Us = United States, Ve = Venezuela.

indicating that the benefit of weighting acoustically prominent segments persists across feature spaces but is amplified when combined with discriminative deep embeddings.

It is important to note that the MFCC-based approach achieved the shortest processing times, demonstrating superior efficiency due to the absence of a neural network for audio feature extraction. However, its hit rate was significantly below the minimum acceptable threshold for a reliable retrieval system. It was also observed that the combination of Perch V2 with the HNSW algorithm achieved the lowest query times and standard deviations compared to Perch V2 with the IMENN, due to its efficient hierarchical data structure and the reduced number of vectors evaluated per query. Regarding the efficiency of fusion techniques, the *Weighted Average pooling* incurs a small time overhead because its calculation is slightly more computationally intensive.

The data employed for the use case is summarized in Table 2, while Figure 3 illustrates the system's practical capability to retrieve acoustically similar recordings using the *Weighted Average Pooling* strategy and Perch V2 embeddings. Following the Stage 2 configuration, one query recording per species was used. The system successfully retrieved matching entries from the vector database, confirming its effectiveness in large-scale retrieval tasks. Furthermore, integrating geographic metadata via the `geopy` library enabled the visualization of species' occurrences, providing a clear spatial representation of their presence across diverse regions.

These results highlight the robustness of the Perch V2 embeddings, which generalize effectively to taxonomic groups absent from its training distribution. Sporadic vocalizations from mammals such as *Alouatta seniculus* and amphibians such as *Andinobates opisthomelas*, as well as non-vocal signals like the stridulation of *Orophus conspersus*,

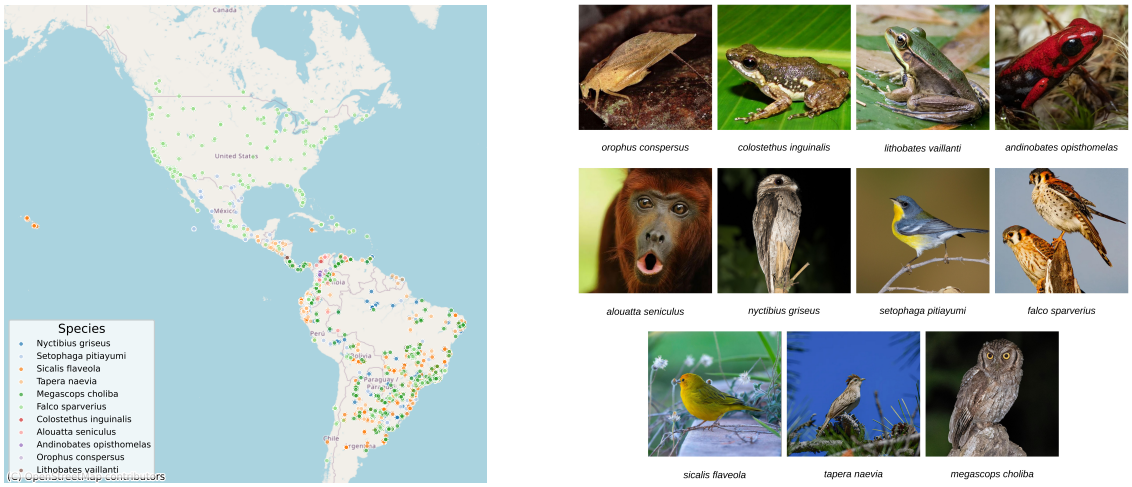

Figure 3. Spatial and taxonomic overview of the Stage 2 use case: (left) geographic distribution map showing the locations of recordings retrieved for the 11 selected species; (right) a visual mosaic illustrating representative individuals of each species used in the retrieval queries.

were retrieved successfully, supporting the hypothesis that weighting segments by RMS energy mitigates the effect of silent intervals and background interference.

The system's utility for biodiversity monitoring and dataset management is exemplified by the retrieval of several key species. For *Colostethus inguinalis* and *Lithobates vaillanti*, two records were identified from the Xeno-Canto (XC) repository [Vellinga, W.P. and Planqué, R., 2025], located in Colombia and Nicaragua, respectively. Similarly, *Orophus conspersus* was identified in Colombia through four records of the insect. Notably, *Andinobates opisthomelas*, categorized as vulnerable, was recovered exclusively from iNaturalist community [2025] (iNat), strictly located within Colombia, demonstrating the system's precision in tracking species of conservation concern in endemic regions.

For larger datasets, the system retrieved bird species with extensive acoustic footprints, such as *Falco sparverius* with 312 recordings and *Megascops choliba* with 401 recordings across more than ten countries in North, Central, and South America, sourced from the XC, iNat, and CSA repositories [Murillo Bedoya et al., 2021]. Additionally, the widespread identification of species such as *Tapera naevia* and *Sicalis flaveola* demonstrated broad distributions spanning the Americas. The resulting spatial distributions revealed high species concentrations in Colombia and Brazil regions encompassing Andean and Amazonian ecosystems renowned for their biological richness and endemism. These findings demonstrate that the proposed retrieval capability, driven by metadata-enriched search, provides valuable insights for ecological studies, species distribution modeling, and the assessment of regional ecological connectivity.

7. Conclusion

This study presented a robust framework for the efficient management, storage, and retrieval of large bioacoustic datasets by integrating deep embeddings, vector fusion strategies, and optimized search algorithms. Our experiments demonstrate that the *Weighted*

Average Pooling technique, combined with a vector database architecture, significantly enhances retrieval efficacy at both the species and recording levels, consistently matching query inputs with their corresponding database entries.

A key finding is the system's ability to generalize across diverse taxonomic groups, including Aves, Mammalia, Amphibia and Insecta. Despite Perch V2 model being pre-trained exclusively on avian vocalizations, this model proved effective for non-avian species, reinforcing the framework's applicability in heterogeneous ecoacoustic contexts. Furthermore, the integration of metadata enabled the visualization of species' distributions, highlighting critical spatial patterns in high-biodiversity regions such as Colombia and Brazil. This capability directly supports ecological research and informed decision-making for habitat identification and conservation.

While the current approach successfully mitigates data challenges through feature fusion, limitations remain regarding class imbalance and the lack of domain-specific embedding adaptation. Future work will focus on implementing domain adaptation techniques, evaluating the system in even broader taxonomic contexts, and validating the proposed methodology against the comprehensive benchmark established by Hamer et al. [2023]. Overall, this work consolidates a scalable and generalizable solution for large-scale acoustic similarity retrieval, providing a robust tool for autonomous biodiversity monitoring and evidence-based conservation.

Acknowledgments

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (AUXPE-CAPES-PROEX), Financing Code 001, and by the Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM, through the PDPG/CAPES and POSGRAD 2026/2027 projects. JGC thanks OpenAI for its support through the partnership established with its representative, Nicolas Robinson Andrade.

References

M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146:3590–3628, 2019. doi: 10.1121/1.5133944.

J. Bjorck, B. H. Rappazzo, D. Chen, R. Bernstein, P. H. Wrege, and C. P. Gomes. Automatic Detection and Compression for Passive Acoustic Monitoring of the African Forest Elephant. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 476–484, 2019. doi: 10.1609/aaai.v33i01.3301476.

Y. Chen, Z. Lai, Y. Ding, K. Lin, and W. K. Wong. Deep supervised hashing with anchor graph. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9796–9804, 2019.

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

J. Hamer, E. Triantafillou, B. van Merriënboer, T. Denton, V. Dumoulin, S. Kahl, and H. Klinck. Birb: A generalization benchmark for information retrieval in bioacoustics. *Preprint under review*, 2023. URL `https://arxiv.org/abs/2312.07439`.

iNaturalist community. iNaturalist – Citizen science platform for biodiversity observations. Online at `https://www.inaturalist.org`, 2025.

A. Jati and D. Emmanouilidou. Supervised deep hashing for efficient audio event retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4497–4501, 2020. doi: 10.1109/ICASSP40776.2020.9053766.

H. Klinck, J. S. Cañas, M. Demkin, S. Dane, S. Kahl, and T. Denton. Birdclef+ 2025. https://kaggle.com/competitions/birdclef-2025, 2025. Kaggle.

O. Krauss, M. Balbino, and C. Nobre. Evaluation of methods of counterfactual explanation - a qualitative and quantitative analysis. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*. SBC, 2023. doi: 10.5753/kdmile.2023.232932.

Y. Liang, S. Zhang, L. K. Li, and X. Wang. Unleashing the full potential of product quantization for large-scale image retrieval. *Advances in Neural Information Processing Systems*, pages 61712–61724, 2023.

Y. A. Malkov and D. A. Yashunin. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

D. Murillo Bedoya, A. Buitrago-Cardona, O. Acevedo-Charry, and J. M. Ochoa-Quintero. Colección de sonidos ambientales mauricio Álvarez-rebolledo (iavh-csa). Instituto Humboldt (Colombia), 2021. URL https://i2d.humboldt.org.co/ceiba/resource.do?r=bancosonidos.

B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti. Soundscape Ecology: The Science of Sound in the Landscape. *BioScience*, 61:203–216, 2011. doi: 10.1525/bio.2011.61.3.6.

M. Ravanelli, B. Elizalde, K. Ni, and G. Friedland. Audio concept classification with hierarchical deep neural networks. In *22nd European Signal Processing Conference (EUSIPCO)*, pages 606–610, 2014.

M. Sert and A. M. Başbuğ. Combining acoustic and semantic similarity for acoustic scene retrieval. In *2019 IEEE International Symposium on Multimedia (ISM)*, 2019. doi: 10.1109/ISM46123.2019.00036.

The IUCN Red List of Threatened Species. Summary statistics – the iucn red list of threatened species, July 2025. URL https://www.iucnredlist.org/resources/summary-statistics.

I. Tolkova. Feature representations for conservation bioacoustics: Review and discussion. *Harvard University*, 2019.

B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton. Perch 2.0: The bittern lesson for bioacoustics. arXiv, 2025. URL https://arxiv.org/abs/2508.04665s.

Vellinga, W.P. and Planqué, R. Xeno-canto – Bird sounds from around the world. GBIF Occurrence Dataset, 2025. URL https://www.gbif.org/dataset/b1047888-ae52-4179-9dd5-5448ea342a24.

A. L.-C. Wang. An industrial strength audio search algorithm. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, 2003.

G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias. Segmentation, indexing, and retrieval for environmental and natural sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):688–707, 2010. doi: 10.1109/TASL.2010.2041384.

J. Xie and M. Zhu. Acoustic classification of bird species using an early fusion of deep features. *Birds*, page 11, 2023. doi: 10.3390/birds4010011. URL https://www.mdpi.com/2673-6004/4/1/11.