# Improving Soundscape Retrieval for Bioacoustic Monitoring: An Analysis of Fusion Techniques with Pre-trained Embeddings

Andrés D. Peralta*
andres@icomp.ufam.edu.br
Federal University of Amazonas
Manaus , AM, Brazil

Eulanda Miranda dos Santos
emsantos@icomp.ufam.edu.br
Federal University of Amazonas
Manaus , AM, Brazil

Marcelo Gordo
mgordo@ufam.edu.br
Federal University of Amazonas
Manaus , AM, Brazil

Jie Xie
xiej8734@gmail.com
Nanji Normal University
Nanjing , Jiangsu, China

Juan G. Colonna
juancolonna@icomp.ufam.edu.br
Federal University of Amazonas
Manaus , AM, Brazil
Victoria University of Wellington
Wellington, New Zealand

## ABSTRACT

The retrieval of similar soundscapes is essential for bioacoustic and ecoacoustic monitoring, yet it remains challenging due to the large volume of unlabeled data, environmental noise, and the complexity of acoustic scenes. To overcome the limitations of traditional, feature-based methods, this study proposes an efficient system that integrates embeddings extracted from a pretrained deep learning model, combined with a noise reduction technique and feature fusion strategies within a vector database to enable similarity-based retrieval. We evaluated the system using bird, amphibian, and mammal recordings across four experimental methodologies, including a use case focused on endangered species. Results show that embedding vectors consistently outperform traditional MFCC (Mel-frequency cepstral coefficients) features in capturing acoustic similarity, and that approximate search algorithms (HNSW) significantly improve both retrieval precision and query efficiency. Additionally, the system effectively retrieves recordings of the critically endangered species *Crax alberti* and maps their geographic distribution, highlighting its potential for conservation planning and early-warning monitoring.

## KEYWORDS

Bioacoustic, Deep Learning, Pre-trained models, Vector database

## 1 INTRODUCTION

Soundscapes comprise natural (biophony and geophony) and anthropogenic (anthropophony) sounds [5]. Their analysis is fundamental in acoustic ecology and bioacoustics [22], especially with the rise of low-cost passive sensors and advances in machine learning [2]. However, soundscape analysis faces challenges such as overlapping sources, adverse noise conditions, and heterogeneous recording devices [23, 24].

In this context, we propose a soundscape retrieval system based on acoustic similarity, which integrates feature fusion techniques as its main contribution. The system leverages embedding vectors extracted from the pretrained Perch model [7], applies advanced feature fusion strategies, and evaluates their performance against traditional MFCC representations. All vectors are stored in a vector database (VectorDB) [14], enabling efficient similarity-based searches on unlabeled acoustic queries. The main goal is to fuse features to retrieve acoustically similar soundscapes from any geographic region using vector-based queries, thus supporting large-scale dataset organization and automated biodiversity monitoring.

Although embedding models and vector databases have been widely used in other domains, such as music similarity retrieval [6, 33], their application in ecoacoustic analysis and bioacoustic similarity retrieval remains relatively underexplored. This study represents an effort to apply vector-based bioacoustic similarity retrieval to a large-scale database of recordings, using Perch and VectorDB in combination. Additionally, we present a systematic comparison between approximate (HNSW) and exact (IMENN) retrieval algorithms, providing a rigorous evaluation of system performance. Perch has also been validated on various bioacoustic tasks, including species classification and cross-domain transfer learning [30], and its open-source implementation promotes scientific integration and reproducibility [25].

Retrieving acoustically similar soundscapes poses significant challenges due to signal variability and large volumes of unlabeled data, making manual annotation time-consuming and often impractical [1, 15]. Vector-based retrieval systems enable efficient querying of acoustic datasets, facilitating the discovery of similar sounds along with their metadata [29]. Traditional hashing algorithms, such as MD5 or SHA256, are ineffective for audio due to their sensitivity to small signal variations [37], while representations like MFCCs project acoustic data into lower-dimensional spaces but remain limited for similarity retrieval tasks [20, 35]. Additionally, ecoacoustic recordings often include unwanted background noise that can hinder analysis [34]. In our study, we applied the Noise Reduce (NR) algorithm, which analysis the noise using the Short-Time Fourier Transform (STFT). The NR algorithm subtracts the average spectral energy per frequency band, effectively attenuating background noise [26].

[*] Both authors contributed equally to this research.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

The remainder of this paper is organized as follows. We formulate the problem in Section 2 and detail the evaluation metrics in Section 3. Related work is reviewed in Section 4. Our methodology is presented in Section 5, followed by the experimental results in Section 6 and our conclusions in Section 7.

## 2 PROBLEM DESCRIPTION

To extract temporal features from raw audio recordings, we model each signal as a one-dimensional time series $s(t) \in \mathbb{R}^T$, 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 $\Delta = 5$ seconds. 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 $s_i \in \mathbb{R}^N$ denote the $i$-th segment of the signal, with $i = 1, \ldots, M$. Each segment is processed by a feature extraction function $f : \mathbb{R}^N \to \mathbb{R}^{1280}$, which produces a fixed-size real-valued feature vector $x_i = f(s_i)$, where $x_i \in \mathbb{R}^{1280}$. In our work, the function $f(\cdot)$ is parameterized by the Perch model [7], a bioacoustic embedding extractor based on the EfficientNet-B1 convolutional neural network architecture. The full audio recording is thus represented by a feature matrix $\mathbf{X} \in \mathbb{R}^{1280 \times M}$, constructed by stacking the feature vectors column-wise $\mathbf{X} = [x_1 \ x_2 \ \ldots \ x_M]$.

Since the input audio duration is variable, the number of columns $M$ in matrix $\mathbf{X}$ also varies across recordings. Thus, we face the challenge to obtain a fixed-length representation suitable for indexing and comparison. Hence, we apply a fusion technique over the time dimension (columns) of $\mathbf{X}$. This produces a global embedding $\bar{x} \in \mathbb{R}^{1280}$, which summarizes the full audio recording as $\bar{x} = \text{Fusion}(\mathbf{X}) \in \mathbb{R}^{1280}$. This step enables consistent downstream processing regardless of the original signal length.

As illustrated in Figure 1, audio signals are preprocessed and transformed into fixed-length embeddings vectors using a pretrained model, which allows an efficient recovery based on similarity. The fusion techniques evaluated in this work are detailed in Section 5.3, including Average Pooling, Weighted Average Pooling, Sum Pooling, and Max Pooling.

## 3 RETRIEVAL ALGORITHMS AND EVALUATION METRICS

The *Hierarchical Navigable Small World* (HNSW) algorithm [19] is used for approximate $k$-nearest neighbors (k-NN) search based on graph structures. This algorithm extends the classical k-NN method by introducing hierarchical graph layers to enable efficient approximate searches. The process begins by selecting an initial node in the graph and building connections to its nearest neighbors based on embedding vector similarities. Nodes are distributed across multiple layers: the upper levels facilitate a quick coarse search, while the lower levels refine the final results. To answer a query, the algorithm starts at a high-level node and traverses down the graph until it reaches a lower level or a maximum number of nodes has been explored. Moreover, 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. The vectors with the highest similarity scores are then returned as the result set.

### 3.1 Evaluation Metrics

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 [16]:

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

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. A low *H@k* value indicates 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 growing impact of machine learning in bioacoustics has driven research into the classification and detection of acoustic events [3, 8]. In [28], the authors proposed an audio scene retrieval system that uses a CNN-GRU model to extract acoustic features, employs the Path Similarity method of WordNet for label inference, and utilizes Mel vectors as spectral descriptors.

In [13], the authors evaluated *Deep Hashing* for efficient retrieval of acoustic events. Pretrained models such as VGGish and TL-Weak [10, 17] were used, alongside a semi-supervised architecture that generates low-dimensional embeddings while optimizing hash codes. The system was trained on the DCASE and ESC-50 datasets, and performance was assessed using the Top-1 metric. Their approach, based on deep hashing and product quantization modules originally designed for image retrieval, showed promising results using non-exhaustive search algorithms

Ghani et al. [7] explored the use of embeddings extracted from pretrained bioacoustic classifiers for transfer learning and novel class recognition tasks. Specifically, they evaluated the ability of embeddings to represent and discriminate new bioacoustic classes not included during training, using datasets comprising bird songs, bat calls, marine mammals, and amphibians. The results suggested that embeddings extracted from bird-trained models outperform those trained on general audio data.

These previous approaches face significant limitations. Several methods, such as deep supervised hashing [4] and product quantization based retrieval [18], are computationally intensive due to the complexity of their training and inference stages, which limits their applicability in large-scale or real-time ecoacoustic monitoring. Additionally, many existing audio representations rely on low-level acoustic features such as frequency, amplitude, and duration which, although useful, may fail to capture the semantic richness of complex acoustic scenes. While vector-based retrieval has been widely adopted in music and entertainment domains [6, 33], its systematic application in ecoacoustics remains limited.

Recent studies have explored feature fusion techniques to improve bioacoustic classification. In [36], the authors proposed early fusion of features extracted from multiple convolutional neural networks to enhance bird species classification performance. Similarly, the review presented in [31] analysis the importance of selecting relevant feature representations to address challenges in conservation bioacoustics, highlighting this choice as a promising strategy for improving generalization. These studies suggest that feature fusion can help capture both global and local acoustic patterns, making it a valuable component in the design of bioacoustic systems.

Additionally, Hamer et al. [9] introduced BIRB, a benchmark designed to evaluate the robustness of vocalization retrieval models against various generalization challenges such as domain shift, novel classes, and few-shot scenarios, using both focal recordings from Xeno-Canto and annotated passive soundscape recordings [9]. BIRB encompasses a diverse set of real-world conditions (different geographic regions, imbalanced class distributions, and recording quality variations) and proposes a baseline retrieval system based on embeddings and centroid search. This work underscores the need to address distributional challenges holistically and provides a framework for comparing transfer learning and domain adaptation approaches in real bioacoustic contexts.

Although recent studies have successfully leveraged embedding vectors for classification and acoustic event identification tasks, we chose not to adopt those models as baselines in this study. This decision stems from their reliance on supervised training, annotated data, or controlled acoustic environments, which may not adequately reflect the complexity of ecoacoustic scenarios. Instead, we adopted classical MFCCs as our baseline, as they constitute a widely used reference for similarity-based audio retrieval.

# 5 MATERIALS AND METHODS

## 5.1 Datasets

This study used the `BirdCLEF+ 2025` dataset, a multitaxonomic bioacoustic collection compiled from various geographic locations and designed for biodiversity monitoring. The dataset comprises a total of 28,564 audio recordings spanning 206 species, including birds, amphibians, mammals, and insects, stored in `.ogg` format with variable sampling rates.

This dataset gathers recordings from three main sources: Xeno-Canto (XC) [32], iNaturalist (iNat) [11], and the Colombian Sound Archive (CSA) [21], which contribute a diverse set of acoustic events. Each recording is accompanied by extensive metadata fields, such as *primary_label*, *type*, *collection*, *scientific_name*, *common_name*, *location* (latitude and longitude), and *author*, enabling detailed and contextualized analysis.

All recordings were preprocessed using the Noise Reduce (NR) algorithm [27], which applies spectral gating to suppress background noise and enhance signal clarity. In the first experiment, the recordings were segmented into non-overlapping 5-second windows to standardize temporal granularity and facilitate uniform embedding extraction. In the remaining experiments, the full-length recordings were used. The distribution of recordings between the vector database and the queries varies depending on the experiment and is explicitly described in Section 5.4.

## 5.2 Proposed Method

Given that bioacoustic recordings often exhibit variable durations and heterogeneous acoustic content, we hypothesize that applying feature fusion strategies can better capture the overall acoustic distribution of each recording. This would allow the system to generalize more effectively across different environments and temporal scales. While previous studies have explored certain forms of feature fusion [36] , our proposal introduces and systematically compares four distinct fusion strategies: average pooling, weighted average pooling, sum pooling, and max pooling. These techniques aim to condense the set of embeddings extracted from each segment into a single representative vector per recording, thus facilitating similarity-based retrieval.

Our method consists of four main stages, which are described below and illustrated in Figure 1.

- **Data Preprocessing:** All audio recordings were resampled to 32kHz and the Noise Reduce[1] algorithm [27] was applied to suppress background noise and enhance signal clarity.
- **Segmentation and Embedding Extraction:** Each recording was segmented into consecutive 5-second windows. Each segment was then processed using the pretrained Perch model to extract 1.280-dimensional embedding vectors.
- **Feature Fusion:** The embeddings extracted from each recording were aggregated into a single representative vector using one of the proposed fusion strategies: average pooling, weighted average pooling, sum pooling, or max pooling.
- **Vector Database Indexing:** The fused vectors, along with their associated metadata, were stored in a vector database (VectorDB[2]), which supports similarity search using the HNSW and IMENN algorithms.

The Perch[3] Deep Learning model is based on the EfficientNet-B1 architecture and was originally trained on bird vocalizations from the Xeno-Canto dataset, sampled at 32kHz. It is optimized to balance performance and computational efficiency, making it well-suited for scalable processing of passive acoustic monitoring data.

In this context, similarity refers to audio segments that share acoustic characteristics, either because they belong to the same species or originate from the same environment or recording session. As a baseline, we implemented traditional MFCCs with 40 coefficients per segment. The MFCCs were extracted using a 1024-point FFT with a hop length of 512 samples and Hann windowing. The same 5-second segments used for embedding extraction were used to compute MFCCs, ensuring fair comparison. Further details on vector merging and evaluation configurations are provided in sections 5.3 and 5.4

[1] https://zenodo.org/records/3243139
[2] https://github.com/jina-ai/vectordb
[3] https://github.com/google-research/perch

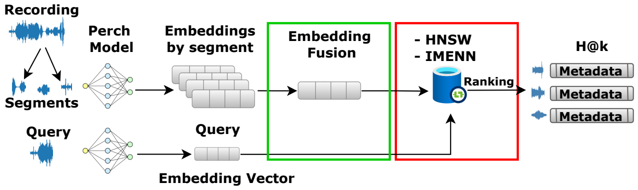

**Figure 1: Proposed method for bioacoustic similarity retrieval of soundscapes.**

## 5.3 Feature Fusion Techniques

To represent each recording with a fixed-size vector, we apply several pooling aggregation techniques described below.

**Average Pooling:** This technique consists of computing the arithmetic mean over the columns of the feature matrix $\mathbf{X} \in \mathbb{R}^{1280\times M}$. This operation yields a single global embedding that summarizes the entire audio signal, providing robustness to isolated acoustic events and mitigating the influence of outlier segments.

The resulting vector representation for a recording is defined as:

$$\bar{x}_{\text{avg}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{:,i} \in \mathbb{R}^{1280}$$

where $\mathbf{X}_{:,i}$ denotes the $i$-th column of matrix $\mathbf{X}$, corresponding to the embedding vector of the $i$-th audio segment. The resulting vectors $\bar{x}_{\text{avg}}$ are stored in the VectorDB and query and recovery using the IMENN and HNSW algorithms.

**Weighted Average Pooling:** In this technique, we incorporate a relevance weight for each segment based on its RSM (Root Mean Square) energy, assigning greater influence to acoustically prominent events such as clear vocalizations. This weighted aggregation emphasizes informative segments while reducing the contribution of low-energy or noisy parts.

Given the feature matrix $\mathbf{X} \in \mathbb{R}^{1280\times M}$, where each column $\mathbf{X}_{:,i}$ represents the embedding of the $i$-th segment, and a corresponding weight vector $\mathbf{w} = [w_1, w_2, \ldots, w_M] \in \mathbb{R}^M$ with $w_i \geq 0$, the weighted average pooled vector is computed as:

$$\bar{x}_{\text{wavg}} = \frac{1}{\sum_{i=1}^{M} w_i}\sum_{i=1}^{M} w_i \cdot \mathbf{X}_{:,i} \in \mathbb{R}^{1280}.$$

This strategy generates a global embedding $\bar{x}_{\text{wavg}}$ that prioritizes segments with high spectral salience. As with average pooling, the resulting vectors are stored in the VectorDB for evaluation with the IMENN and HNSW retrieval algorithms.

**Sum Pooling:**

In this technique, the global embedding is computed by summing all column vectors of the feature matrix $\mathbf{X} \in \mathbb{R}^{1280\times M}$, which represent the individual segments of a recording. The resulting vector retains the same dimensionality and captures cumulative acoustic information across time:

$$\bar{x}_{\text{sum}} = \sum_{i=1}^{M}\mathbf{X}_{:,i} \in \mathbb{R}^{1280}$$

This approach emphasizes aggregate patterns and the joint contribution of all segments while maintaining computational simplicity. However, it may be sensitive to outlier segments or regions with unusually high energy if not further normalized.

**Max Pooling:** The *max pooling* technique constructs the global embedding by selecting the maximum value along each feature dimension from all columns of the feature matrix $\mathbf{X} \in \mathbb{R}^{1280\times M}$. This highlights the most prominent acoustic features, regardless of when they occur in the recording:

$$\bar{x}_{\text{max}} = \max_{i=1,\ldots,M} \mathbf{X}_{:,i} \in \mathbb{R}^{1280}$$

where the max operation is applied element-wise across columns (i.e., across time) for each of the 1280 dimensions. This method accentuates dominant acoustic events, such as intense vocalizations, although it may be susceptible to noise.

All the described fusion techniques aim to capture both global and local acoustic patterns from the recordings, enhance inter-class discrimination, and provide robustness in retrieval tasks under varying acoustic conditions. Regardless of the fusion strategy employed, the resulting vector $\bar{x}$ has a fixed dimensionality of 1280, ensuring compatibility and computational efficiency within the VectorDB framework. Each vector is also linked to detailed metadata, as described in Section 5.1.

For baseline comparison, we also extracted 128-dimensional MFCC feature vectors from each 5-second audio segment, resulting in matrices $\mathbf{X} \in \mathbb{R}^{128\times M}$. The same data fusion techniques were applied to these MFCC matrices, producing aggregated vectors $\bar{x} \in \mathbb{R}^{128}$, allowing fair comparison between classical and deep embedding-based representations.

It is worth noting that the application of these fusion techniques varied depending on the experiment. In Experiment 1, fusion was applied to both the database and queries. In Experiment 2, it was used exclusively in the database, keeping the query unaggregated. Conversely, in Experiment 3, fusion was applied only to the queries, with the database remaining segmented. Finally, Experiment 4 adopted the same configuration as Experiment 2, which achieved the best performance, to transfer it to the final full indexing scenario.

## 5.4 Evaluation Methodologies

The evaluation was structured into two experimental stages. In stage 1, we compared three different retrieval configurations to identify the most effective feature fusion technique, as illustrated in Figure 2.

In the first experimental configuration, a standard 70%-30% split was applied at the recording level for each species. This means that 30% of the recordings from each species were used exclusively as queries, while the remaining 70% were assigned to the database. This setup ensured that there was no temporal or source overlap between the query and database sets, thereby eliminating the risk of information leakage. Retrieval performance was evaluated at the species level, allowing us to assess the system's ability to identify acoustically similar segments across different recordings of the same species. Additionally, this configuration was used to systematically compare how different feature fusion strategies affect retrieval accuracy and to evaluate how the traditional partitioning methodology—commonly adopted in machine learning—combined with various fusion techniques, influences the system's ability to retrieve relevant audio segments.

In the second experimental configuration, the segment with the highest energy from each recording was selected as the query, while all segments from the same recording were stored in the vector database, intentionally introducing temporal and source overlap. Retrieval performance was evaluated both by species and by recording. This experimental setup assesses whether a prominent acoustic event, such as a clear vocalization, can serve as an effective query to retrieve related segments from similar species or recordings in noisy environments. Moreover, this configuration enables the evaluation of how different feature fusion strategies affect retrieval precision when queries are based on salient acoustic events.

In the third experimental configuration, the segment with the highest energy from each recording was stored in the vector database, while the remaining segments from the same recording were used as queries. As in the previous configuration, this strategy introduces temporal and source overlap between the query and database sets, reflecting a more realistic continuous monitoring scenario in which acoustic events often share partial information. Retrieval was evaluated at both the species and recording levels. This configuration enables the analysis of whether a prominent acoustic event can be effectively retrieved using less prominent related segments.

Additionally, for each experiment, individual query times were measured, considering exclusively the retrieval phase in VectorDB and excluding the time dedicated to embedding extraction or fusion. Times were measured for each query separately and then averaged to obtain the mean time and standard deviation reported in the results tables. In Experiment 1, 130.245 segments were used for the database and 57.594 for queries; in Experiment 2, 187.839 segments in the database and 26.250 for queries; and in Experiment 3, 26.250 in the database and 187.839 for queries.

In Stage 2, with the goal of demonstrating the applicability of the system, a use case was designed focusing on the functional interaction between feature researchers dedicated to bioacoustics, species conservation, and wildlife monitoring, and the proposed retrieval system. In this scenario, the user uploads an unlabeled reference recording, and the system returns a set of acoustically similar recordings, along with their respective metadata: species, approximate geographic location, and source. This functionality, supported by pretrained embeddings, feature fusion techniques, and vector-based search, enables faster acoustic data analysis, reduces manual review workload, and will support decision-making in biodiversity monitoring and conservation contexts.

Using the second experimental configuration defined in Stage 1, along with the weighted average pooling fusion technique which yielded the best results the system was configured to retrieve similar recordings grouped by species within the vector database. This final experiment focused on thirteen representative species from the BirdCLEF+ 2025 dataset: *Elaenia flavogaster*, *Penelope purpurascens*, *Megarynchus pitangua*, *Andinobates opisthomelas*, *Pyrilia pyrilia*, *Panthera onca*, *Alouatta seniculus*, *Bradypus variegatus*, *Colostethus inguinalis*, *Cerdocyon thous*, *Allobates niputidea*, *Lontra longicaudis*, and *Crax alberti*. These species were selected based on their conservation status and population decline risk (according to Red List categories), determined by cross-referencing the metadata of the recordings with the IUCN Red List [12].

For each retrieved recording, key metadata such as geographic location (latitude and longitude) and source (XC, iNat, or CSA) were analyzed. Additionally, the geopy[4] library was used in combination with the Nominatim service to obtain the approximate location associated with each acoustic recording. A 2D geographic map was generated to visualize the spatial distribution of the species and highlight potential spatial patterns. The aim of this approach is to demonstrate the system's potential to support ecological research and conservation actions, enabling biologists to monitor distribution changes and identify critical habitats. The combination of the map and metadata table facilitates rapid, informed ecological interpretation for decision-making.

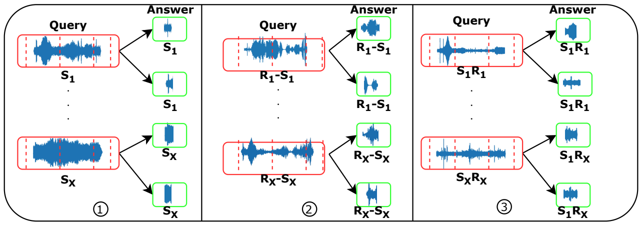

**Figure 2: Methodological approaches: (1) 70%-30% split at recording level; (2) highest-energy segment as query; (3) highest-energy segment as database, remaining segments as queries.**

# 6 RESULTS

## 6.1 Results of Experiment 1

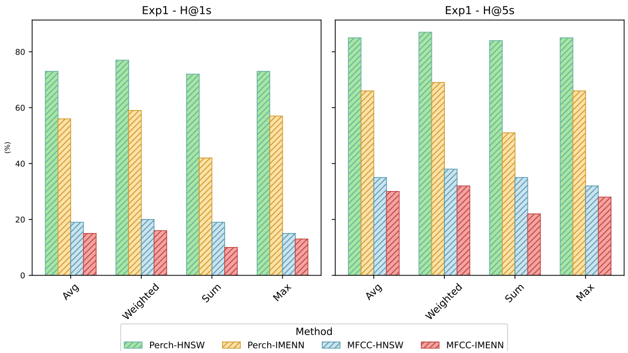

**Figure 3: Results of the comparison of fusion techniques and methods for recovery at the species level.**

The results of Experiment 1, presented in Table 1 and Figure 3 , confirm that the embeddings extracted using the pretrained Perch model greatly outperformed traditional MFCC representations, especially after applying the noise reduction filter. The improvement observed in precision metrics for species-level retrieval ($H@1_S$,

[4] https://pypi.org/project/geopy/

**Table 1: Experiment 1 results: species-level retrieval comparison using embeddings and MFCCs. H@$_s$ denotes species-level evaluation. $t \pm \sigma$ indicates the average query time (in milliseconds) and its standard deviation.**

| Perch | | | | | | |
|---|---|---|---|---|---|---|
| | HNSW | | | IMMEN | | |
| Technique | H@$1_s$ | H@$5_s$ | t±$\sigma$ | H@$1_s$ | H@$5_s$ | t±$\sigma$ |
| Avg Pool | 0.73 | 0.85 | 19 ± 0.17 | 0.56 | 0.66 | 58 ± 6.86 |
| Weighted Avg | 0.77 | 0.87 | 21 ± 0.22 | 0.59 | 0.69 | 58 ± 6.90 |
| Sum | 0.72 | 0.84 | 16 ± 0.20 | 0.42 | 0.51 | 59 ± 6.79 |
| Max Pool | 0.73 | 0.85 | 17 ± 0.19 | 0.57 | 0.66 | 59 ± 6.79 |
| MFCCs | | | | | | |
| Avg Pool | 0.19 | 0.35 | 5 ± 0.05 | 0.15 | 0.30 | 8 ± 3.37 |
| Weighted Avg | 0.20 | 0.38 | 6 ± 0.06 | 0.16 | 0.32 | 9 ± 3.41 |
| Sum | 0.19 | 0.35 | 4 ± 0.05 | 0.10 | 0.22 | 7 ± 3.62 |
| Max Pool | 0.15 | 0.32 | 5 ± 0.05 | 0.13 | 0.28 | 8 ± 4.22 |

$H@5_s$) underscores the importance of preprocessing, given the high level of environmental interference and the variability introduced by different devices and recording locations. Specifically, Perch embeddings combined with the *Weighted Average pooling* technique and the HNSW algorithm achieved the highest accuracy values, reaching $H@1_s = 0.77$ and $H@5_s = 0.87$.

It was also observed that the HNSW algorithm consistently achieved the lowest query times and standard deviations, due to its efficient hierarchical structure and the lower number of vectors evaluated per query. In summary, the experiment confirms that combining perceptually rich embeddings, effective feature fusion strategies, and approximate search algorithms can enhance both accuracy and efficiency in acoustic similarity retrieval.

## 6.2 Results of Experiment 2

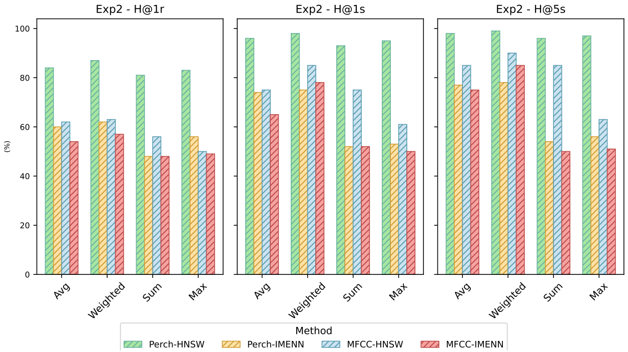

**Figure 4: Results of the fusion technique by querying the highest energy segments by recording and species.**

The results of Experiment 2, shown in Table 2 and Figure 4, confirm the effectiveness of using the highest-energy segment from each recording as the query while storing the full recording in the vector database. Retrieval was evaluated at both the species level ($H@1_s, H@5_s$) and the recording level ($H@1_r$), allowing us to assess the system's ability to identify related species and exact matches between recordings.

Compared to Experiment 1, a significant improvement in acoustic retrieval was observed when using prominent acoustic events as queries. This result reinforces the importance of selecting segments with high information density and demonstrates how fusion techniques—especially *Weighted Average pooling*—can enhance retrieval quality even when using traditional representations such as MFCCs. Despite the improvement in MFCC-based retrieval, the embeddings generated with the Perch model continued to deliver the best performance, achieving $H@1_s = 0.98$, $H@5_s = 0.99$, and $H@1_r = 0.87$ when combined with the HNSW algorithm. Regarding query times, the HNSW algorithm consistently exhibited lower response times compared to IMENN, supporting its suitability for high-demand scenarios involving large-scale data and multiple simultaneous queries.

The advantage of using prominent segments—such as clear vocalizations—to initiate queries is clear: this strategy not only improves precision but also enhances the robust identification of species and entire recordings, even under noisy and variable conditions.

## 6.3 Results of Experiment 3

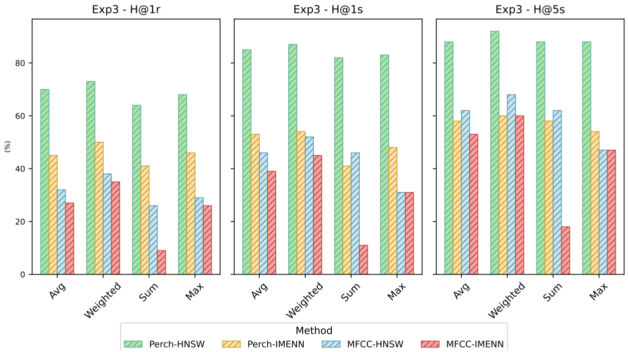

**Figure 5: Results of the fusion technique saving higher energy segments in the database by recording and species.**

The results of Experiment 3, summarized in Table 3 and Figure 5, confirm the consistent performance of the system when using the highest-energy segment of each recording as the stored representation in the database, while the remaining segments were used as individual queries. This configuration allowed us to assess retrieval effectiveness in a more realistic scenario, where different portions of the same recording are compared to verify acoustic similarity.

In this setting, the combination of embeddings generated by the Perch model with the *Weighted Average pooling* technique

**Table 2: Experiment 2 results: query using the highest-energy segment and full database by recording and species.**

| Perch | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HNSW | | | | | | IMMEN | | | | |
| **Technique** | **H@1$_r$** | **t±$\sigma$** | **H@1$_s$** | **H@5$_s$** | **t±$\sigma$** | | **H@1$_r$** | **t±$\sigma$** | **H@1$_s$** | **H@5$_s$** | **t±$\sigma$** |
| Avg Pool | 0.84 | 17 ±$\sigma$ 0.08 | 0.96 | 0.98 | 23 ±$\sigma$ 0.11 | | 0.60 | 43 ±$\sigma$ 3.01 | 0.74 | 0.77 | 62 ±$\sigma$ 6.92 |
| Weighted Avg | 0.87 | 21 ±$\sigma$ 0.12 | 0.98 | 0.99 | 29 ±$\sigma$ 0.17 | | 0.62 | 40 ±$\sigma$ 3.27 | 0.75 | 0.78 | 71 ±$\sigma$ 6.77 |
| Sum | 0.81 | 15 ±$\sigma$ 0.09 | 0.93 | 0.96 | 26 ±$\sigma$ 0.14 | | 0.48 | 34 ±$\sigma$ 2.90 | 0.52 | 0.54 | 68 ±$\sigma$ 6.88 |
| Max Pool | 0.83 | 19 ±$\sigma$ 0.10 | 0.95 | 0.97 | 24 ±$\sigma$ 0.12 | | 0.56 | 38 ±$\sigma$ 3.00 | 0.53 | 0.56 | 63 ±$\sigma$ 7.42 |
| MFCCs | | | | | | | | | | | |
| Avg Pool | 0.62 | 8 ±$\sigma$ 0.10 | 0.75 | 0.85 | 10 ±$\sigma$ 0.04 | | 0.54 | 13 ±$\sigma$ 0.26 | 0.65 | 0.75 | 14 ±$\sigma$ 3.61 |
| Weighted Avg | 0.63 | 11 ±$\sigma$ 0.13 | 0.85 | 0.90 | 13 ±$\sigma$ 0.07 | | 0.57 | 15 ±$\sigma$ 0.29 | 0.78 | 0.85 | 16 ±$\sigma$ 3.33 |
| Sum | 0.56 | 9 ±$\sigma$ 0.08 | 0.75 | 0.85 | 9 ±$\sigma$ 0.06 | | 0.48 | 11 ±$\sigma$ 0.25 | 0.52 | 0.50 | 12 ±$\sigma$ 3.42 |
| Max Pool | 0.50 | 7 ±$\sigma$ 0.07 | 0.61 | 0.63 | 12 ±$\sigma$ 0.04 | | 0.49 | 9 ±$\sigma$ 0.23 | 0.50 | 0.51 | 15 ±$\sigma$ 3.40 |

**Table 3: Results of the fusion technique saving higher energy segments in the database by recording and species.**

| Perch | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HNSW | | | | | | IMENN | | | | |
| **Technique** | **H@1$_r$** | **t±$\sigma$** | **H@1$_s$** | **H@5$_s$** | **t±$\sigma$** | | **H@1$_r$** | **t±$\sigma$** | **H@1$_s$** | **H@5$_s$** | **t±$\sigma$** |
| Avg Pool | 0.70 | 5 ±$\sigma$ 0.21 | 0.85 | 0.88 | 19 ± 0.75 | | 0.45 | 6 ±$\sigma$ 0.33 | 0.53 | 0.58 | 43 ± 6.71 |
| Weighted Avg | 0.73 | 6 ±$\sigma$ 0.29 | 0.87 | 0.92 | 23 ± 0.14 | | 0.50 | 8 ±$\sigma$ 0.35 | 0.54 | 0.60 | 44 ± 6.34 |
| Sum | 0.64 | 6 ±$\sigma$ 0.22 | 0.82 | 0.88 | 18 ± 0.11 | | 0.41 | 7 ±$\sigma$ 0.30 | 0.41 | 0.58 | 48 ± 6.26 |
| Max Pool | 0.68 | 4 ±$\sigma$ 0.23 | 0.83 | 0.88 | 21 ± 0.14 | | 0.46 | 6 ±$\sigma$ 0.31 | 0.48 | 0.54 | 42 ± 7.10 |
| MFCCs | | | | | | | | | | | |
| Avg Pool | 0.32 | 6 ±$\sigma$ 0.35 | 0.46 | 0.62 | 3 ± 0.74 | | 0.27 | 8 ±$\sigma$ 0.36 | 0.39 | 0.53 | 4 ± 3.24 |
| Weighted Avg | 0.38 | 8 ± 0.41 | 0.52 | 0.68 | 4 ± 0.70 | | 0.35 | 10 ± 0.45 | 0.45 | 0.60 | 6 ± 3.38 |
| Sum | 0.26 | 6 ± 0.33 | 0.46 | 0.62 | 2 ± 0.53 | | 0.09 | 7 ± 0.38 | 0.11 | 0.18 | 3 ± 2.84 |
| Max Pool | 0.29 | 7 ± 0.35 | 0.31 | 0.47 | 3 ± 0.55 | | 0.26 | 8 ± 0.37 | 0.31 | 0.47 | 4 ± 3.20 |

and the HNSW algorithm again delivered high retrieval effectiveness, achieving $H@1_s = 0.87$, $H@5_s = 0.92$, and $H@1_r = 0.73$. While MFCCs showed improvements when combined with appropriate fusion techniques, the Perch embeddings continued to outperform, especially in capturing relevant acoustic patterns across intra-recording variations. This experiment recorded the lowest average query times, a direct consequence of the reduced number of queries. Regarding efficiency, the HNSW algorithm maintained the trend of achieving the lowest query times and standard deviations across all evaluated configurations.

## 6.4 Results of Use Case

The use case, illustrated in Table 4 and Figure 6, demonstrates the practical capability of the system to retrieve acoustically similar recordings using the *Weighted Average Pooling* fusion technique and pretrained embedding vectors. In this experiment, one query recording per species was used, following the configuration defined in Stage 2. The system successfully retrieved recordings from the vector database that matched the input recording, confirming its effectiveness in retrieval tasks. Additionally, by combining geographic metadata with the geopy library, it was possible to visualize the locations where each species was recorded, providing a useful representation of their observed presence across different regions.

The system correctly retrieved the corresponding recordings from the vector database, all matching the queried species. For *Lontra longicaudis*, two recordings were identified in Brazil, sourced from the XC and iNat repositories. *Allobates niputidea* was found in Colombia through two recordings registered in CSA. The species *Bradypus variegatus*, *Colostethus inguinalis*, and *Cerdocyon thous* were also located in Colombia and Brazil, based on two recordings extracted from iNat. *Crax alberti*, listed as Critically Endangered, was exclusively retrieved from Colombia via XC. In the case of *Andinobates opisthomelas*, ten recordings were recovered, all from Colombia and sourced from iNat. *Pyrilia pyrilia* had recordings from Colombia and Panama, extracted from XC and iNat.

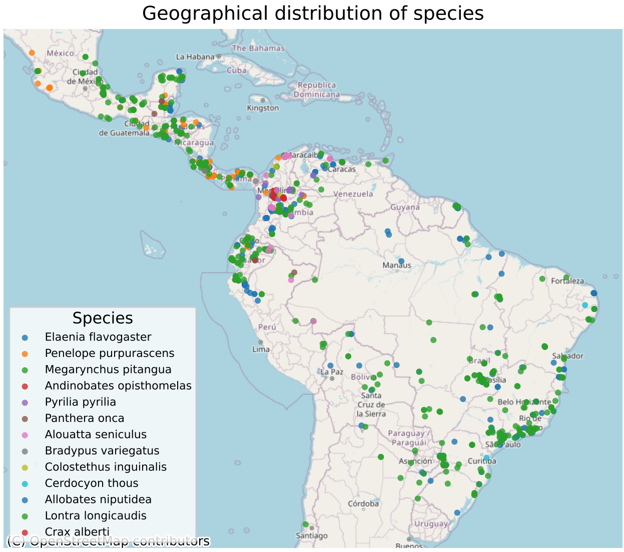

**Figure 6: Results of geographic distribution and retrieval performance of selected species.**

For *Panthera onca*, the system retrieved fifteen recordings distributed across Brazil, Colombia, and Mexico. *Alouatta seniculus* had records in both Brazil and Colombia, while *Penelope purpurascens* was identified in Colombia, Costa Rica, Mexico, Panama, and

**Table 4: Use Case: species, recordings, and geographic distribution.**

| Species | # Recordings | Taxonomic Class | Source | Red List Category | Location |
|---|---|---|---|---|---|
| *Lontra longicaudis* | 2 | Mammalia | XC (1)<br>iNat (1) | Near Threatened | Br (2) |
| *Allobates niputidea* | 2 | Amphibia | CSA (2) | Least Concern | Col (2) |
| *Bradypus variegatus* | 2 | Mammalia | iNat (2) | Least Concern | Br (1), Col (1) |
| *Colostethus inguinalis* | 2 | Amphibia | iNat (2) | Least Concern | Col (2) |
| *Cerdocyon thous* | 2 | Mammalia | iNat (2) | Least Concern | Br (2) |
| *Crax alberti* | 8 | Aves | XC (8) | Critically Endangered | Col (8) |
| *Andinobates opisthomelas* | 10 | Amphibia | iNat (10) | Vulnerable | Col (10) |
| *Pyrilia pyrilia* | 14 | Aves | XC (13)<br>iNat (1) | Near Threatened | Col (12), Pa (2) |
| *Panthera onca* | 15 | Mammalia | XC (11)<br>iNat (4) | Near Threatened | Br (7), Col (7), Mx (1) |
| *Alouatta seniculus* | 23 | Mammalia | XC (1)<br>iNat (22) | Least Concern | Br (7), Col (16) |
| *Penelope purpurascens* | 77 | Aves | XC (61)<br>iNat (16) | Near Threatened | Col (60), CR (6), Mx (6), Pa (4), Ni (1) |
| *Elaenia flavogaster* | 260 | Aves | XC (226)<br>iNat (34) | Least Concern | Ar (8), Bz (10), Bo (6), Br (108), Col (60), CR (12), Ec (16), SV (1), Fr (5), Hn (4), Mx (5), Pa (9), Py (1), Pe (6), Ve (9) |
| *Megarynchus pitangua* | 580 | Aves | XC (339)<br>iNat (169) | Least Concern | Ar (1), Bo (1), Br (302), Col (214), CR (16), Ec (16), SV (1), Mx (10), Pa (11), Pe (5), Ve (3) |

**Location codes:** Ar = Argentina, Bz = Belize, Bo = Bolivia, Br = Brazil, Col = Colombia, CR = Costa Rica, Ec = Ecuador, SV = El Salvador, Fr = France, Hn = Honduras, Mx = Mexico, Ni = Nicaragua, Pa = Panama, Py = Paraguay, Pe = Peru, Ve = Venezuela.

Nicaragua, with recordings from XC and iNat. Finally, *Elaenia flavogaster* and *Megarynchus pitangua*, the species with the highest number of recordings, exhibited a broad geographic distribution across Latin America, with records from more than ten countries, mostly collected from XC and iNat. This retrieval enables the visualization, based on metadata, of the regions where these species were observed, providing valuable information for ecological studies and monitoring strategies.

From a taxonomic perspective, the species retrieved in this use case span three classes: Mammalia, Aves, and Amphibia. Regarding conservation categories defined by the IUCN Red List, the system retrieved recordings of all selected species, including *Crax alberti* (Critically Endangered), *Andinobates opisthomelas* (Vulnerable), and four Near Threatened species: *Lontra longicaudis*, *Pyrilia pyrilia*, *Panthera onca*, and *Penelope purpurascens*. The remaining seven species are classified as Least Concern. Although species were selected based on their vulnerability and limited number of recordings, this use case demonstrates that the system was able to retrieve existing acoustic records from the `BirdCLEF+ 2025` database associated with species from different taxonomic classes. This highlights the system's ability to generalize acoustic retrieval across multiple biological groups, which is crucial for comprehensive biodiversity inventories.

The map generated using the geographic coordinates available in the recording metadata allowed for the visualization of the locations where each species was recorded. As a result, a high concentration of species was observed in Colombia and Brazil—countries that include Andean and Amazonian ecosystems, widely recognized for their biological richness and high endemism. These observations, derived from the retrieval results, can support future ecological studies focused on species distribution or exploring ecological connectivity between regions.

# 7 CONCLUSION

This study proposed an efficient approach for retrieving acoustically similar recordings in unlabeled ecoacoustic databases, integrating deep embeddings, feature fusion techniques, and approximate search algorithms within a vector database. The experiments demonstrated that the *Weighted Average Pooling* technique significantly improved retrieval efficacy at both the species and recording levels.

The use case demonstrated that the system was able to successfully retrieve recordings corresponding to the same species as the query, even when they belonged to different taxonomic classes (Aves, Mammalia, and Amphibia). Although the Perch model was originally trained on bird vocalizations, the system exhibited a capacity to generalize to non-avian species. This taxonomic diversity reinforces its applicability in heterogeneous ecoacoustic contexts, where data from multiple animal groups coexist. Furthermore, the integration of metadata enabled the visualization of species' geographic distributions, revealing concentrations in Colombia and Brazil, regions recognized for their high levels of endemism.

Among the main limitations identified is the imbalance in the number of recordings per species. Furthermore, our approach relies on pre-trained embeddings without domain adaptation. However, the feature fusion approach helped mitigate this issue by enhancing the acoustic representations. Overall, this work consolidates an effective and generalizable system for acoustic similarity retrieval, with potential applications in ecological studies, biodiversity monitoring, and evidence-based conservation decision-making. As future work, we propose evaluating the system in acoustic contexts with greater taxonomic diversity, implementing domain adaptation techniques for the embeddings, and validating our approach within the comprehensive benchmark proposed by [9].

## ACKNOWLEDGMENTS

This work was supported by the Coordenação de Aperfeiçoa-mento 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 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. This work is also supported by the National Natural Science Foundation of China (Grant No. 32371556).

## REFERENCES

[1] Luke Barrington, Antoni Chan, Douglas Turnbull, and Gert Lanckriet. 2007. Audio Information Retrieval using Semantic Similarity. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Vol. 2. II–725–II–728. https://doi.org/10.1109/ICASSP.2007.366338

[2] Michael J. Bianco, Peter Gerstoft, James Traer, Emma Ozanich, Marie A. Roch, Sharon Gannot, and Charles-Alban Deledalle. 2019. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America* 146 (2019), 3590–3628. https://doi.org/10.1121/1.5133944

[3] Johan Bjorck, Brendan H. Rappazzo, Di Chen, Richard Bernstein, Peter H. Wrege, and Carla P. Gomes. 2019. Automatic Detection and Compression for Passive Acoustic Monitoring of the African Forest Elephant. (2019), 476–484. https://doi.org/10.1609/aaai.v33i01.3301476

[4] Yudong Chen, Zhihui Lai, Yujuan Ding, Kaiyi Lin, and Wai Keung Wong. 2019. Deep supervised hashing with anchor graph. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9796–9804.

[5] Dhanunjaya Varma Devalraju and Padmanabhan Rajan. 2022. Multiview Embeddings for Soundscape Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1197–1206. https://doi.org/10.1109/TASLP.2022.3153272

[6] Spotify Engineering. 2020. Approximate Nearest Neighbor Search for Audio Embeddings at Spotify. https://engineering.atspotify.com/2020/07/approximate-nearest-neighbor-search-for-audio-embeddings-at-spotify/ Accessed: 2025-06-26.

[7] Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports* (2023). https://doi.org/10.1038/s41598-023-49989-z

[8] Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. 2023. BEANS: The Benchmark of Animal Sounds. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096686

[9] Jenny Hamer, Eleni Triantafillou, Bart van Merriënboer, Tom Denton, Vincent Dumoulin, Stefan Kahl, and Holger Klinck. 2023. BIRB: A Generalization Benchmark for Information Retrieval in Bioacoustics. *Preprint under review* (2023). https://arxiv.org/abs/2312.07439

[10] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP), 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

[11] iNaturalist community. 2025. iNaturalist – Citizen science platform for biodiversity observations. Online at https://www.inaturalist.org.

[12] IUCN Red List. 2024. The IUCN Red List of Threatened Species. https://www.iucnredlist.org/

[13] Arindam Jati and Dimitra Emmanouilidou. 2020. Supervised Deep Hashing for Efficient Audio Event Retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4497–4501. https://doi.org/10.1109/ICASSP40776.2020.9053766

[14] Jina-Ai. 2023. Jina-ai/vectordb: A Python vector database you just need - no more, no less. https://github.com/jina-ai/vectordb

[15] A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2023. Audio Retrieval With Natural Language Queries: A Benchmark Study. *IEEE Transactions on Multimedia* 25 (2023), 2675–2685. https://doi.org/10.1109/TMM.2022.3149712

[16] Omar Krauss, Marcelo Balbino, and Cristiane Nobre. 2023. Evaluation of methods of counterfactual explanation - A qualitative and quantitative analysis. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*. SBC. https://doi.org/10.5753/kdmile.2023.232932

[17] Anurag Kumar, Maksim Khadkevich, and Christian Fügen. 2018. Knowledge Transfer from Weakly Labeled Audio Using Convolutional Neural Network for Sound Events and Scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 326–330. https://doi.org/10.1109/ICASSP.2018.8462200

[18] Yu Liang, Shiliang Zhang, Li Ken Li, and Xiaoyu Wang. 2023. Unleashing the full potential of product quantization for large-scale image retrieval. *Advances in Neural Information Processing Systems* (2023), 61712–61724.

[19] Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

[20] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *24th European Signal Processing Conference (EUSIPCO)*. IEEE, 1128–1132.

[21] Murillo Bedoya, D. and Buitrago-Cardona, A. and Acevedo-Charry, O. and Ochoa-Quintero, J. M. 2021. Colección de Sonidos Ambientales Mauricio Álvarez-Rebolledo (IAvH-CSA). Instituto Humboldt (Colombia). https://www.humboldt.org.co/

[22] Bryan C. Pijanowski, Luis J. Villanueva-Rivera, Sarah L. Dumyahn, Almo Farina, Bernie L. Krause, Brian M. Napoletano, Stuart H. Gage, and Nadia Pieretti. 2011. Soundscape Ecology: The Science of Sound in the Landscape. *BioScience* 61 (2011), 203–216. https://doi.org/10.1525/bio.2011.61.3.6

[23] Shah Jafor Sadeek Quaderi, Sadia Afrin Labonno, Sadia Mostafa, and Shamim Akhter. 2022. Identify The Beehive Sound Using Deep Learning. *arXiv.org* (2022). https://doi.org/10.48550/arXiv.2209.01374

[24] Mirco Ravanelli, Benjamin Elizalde, Karl Ni, and Gerald Friedland. 2014. Audio concept classification with Hierarchical Deep Neural Networks. In *22nd European Signal Processing Conference (EUSIPCO)*. 606–610.

[25] Google Research. 2023. Perch-Hoplite: A repository for bird sound classification and few-shot learning. https://github.com/google-research/perch-hoplite

[26] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology* 16, 10 (2020).

[27] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology* 16, 10 (2020).

[28] Mustafa Sert and Ahmet Melih Başbuğ. 2019. Combining Acoustic and Semantic Similarity for Acoustic Scene Retrieval. In *2019 IEEE International Symposium on Multimedia (ISM)*. https://doi.org/10.1109/ISM46123.2019.00036

[29] Malcolm Slaney. 2002. Semantic-audio retrieval. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IV–4108–IV–4111. https://doi.org/10.1109/ICASSP.2002.5745561

[30] Kevin Smith, Uzay Ghani, and Juan G. Colonna. 2024. Towards Deep Active Learning in Avian Bioacoustics. In *ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

[31] Irina Tolkova. 2019. Feature Representations for Conservation Bioacoustics: Review and Discussion. *Harvard University* (2019). https://projects.iq.harvard.edu/sites/projects.iq.harvard.edu/files/crcs/files/ai4sg-21_paper_43.pdf

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

[33] Avery Li-Chun Wang. 2003. An Industrial Strength Audio Search Algorithm. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*.

[34] Cheng Wang, Haojin Yang, and Christoph Meinel. 2015. Deep Semantic Mapping for Cross-Modal Retrieval. In *IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. 234–241. https://doi.org/10.1109/ICTAI.2015.45

[35] Gordon Wichern, Jiachen Xue, Harvey Thornburg, Brandon Mechtley, and Andreas Spanias. 2010. Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds. *IEEE Transactions on Audio, Speech, and Language Processing* 3 (2010), 688–707. https://doi.org/10.1109/TASL.2010.2041384

[36] Jie Xie and Mingying. Zhu. 2023. Acoustic Classification of Bird Species Using an Early Fusion of Deep Features. *Birds* (2023), 11. https://doi.org/10.3390/birds4010011

[37] Hanxiao Xu. 2020. Cross-Modal Sound-Image Retrieval Based on Deep Collaborative Hashing. In *5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*. 188–197. https://doi.org/10.1109/ISCTT51595.2020.00041