Acoustic identification of individual animals in the wild

Inês de Almeida Nolasco

Submitted in partial fulfilment of the
requirements of the Degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London

This work was supported by the Engineering and Physical Sciences
Research Council [grant number EP/R513106/1]

2025

Abstract

Traditionally, studies of animal vocal communication have focused on species-level analysis. While this approach has provided valuable insights, it often overlooks the individuality present within populations. Recent advances in technology, combined with novel research perspectives, have highlighted the importance of studying individual-level variation. This shift in focus reveals unique behaviours, ecological roles, and social dynamics that are critical for a deeper understanding of animal populations.

In this thesis, we develop computational methods to advance automatic acoustic identification on individuals (AIID), with the goal of identifying individual animals based on their unique vocal characteristics. The latest advancements in machine learning (ML) inspire the development of robust methodologies design to perform AIID in real-world scenarios. The methods are developed to address key challenges such as generalization across multiple species and open world classification, where neither the species nor the individuals present are fully known.

Unlike traditional species-level bioacoustic studies, this work adopts a novel hierarchical framework, accounting for the taxonomic relationships among species and individuals. This approach not only enhances AIID performance but also provides new insights into vocal recognition mechanisms across the different taxonomic groups.

The thesis first establishes a baseline for multispecies AIID by curating a novel dataset comprising vocalisations annotated at the individual level for both mammals and birds. Classification-based training methods are adopted to first evaluate the benefits of incorporating taxonomic information into the training process. To address the inherent challenges of AIID in the wild, particularly open world classification and the limited availability of annotated data, the thesis further explores alternative training paradigms based on distance learning.

The results demonstrate that integrating hierarchical constraints into AIID frameworks significantly improves identification performance across all taxonomic levels while enhancing generalization to unseen individuals and species. These findings advance the technical capabilities of AIID systems and underscore their potential to drive progress in animal behaviour research and conservation efforts.

Acknowledgements

I would like to thank everyone who was present during the long journey of this thesis. First, this work would have not been possible without the recordings that were generously shared with me. Thank you to the many recordists and annotators. To my many collaborators, thank you for making the work enjoyable and engaging.

To my supervisors, thank you for all the insightful discussions and guidance throughout this process. A very special thank you to Emmanouil and Dan, for your kindness, support, and example.

To all my friends near and far away, that filled this time with excitement and learning. A special thank you to my London family: Aditya, for believing in me with such strength that I cannot help but feel a confidence boost whenever we meet. I'll forever cherish our long talks. Charlotte, my other phD student/mummy, our shared experience made it feel so less lonely, I'm glad we had each other to lean on. Eurico, the PhD is done! I'm so happy to have you in London and thank you for being an example on how to enjoy life to the fullest. V, JT, Benny and Courtney, it is so hard to tell you what our little group meant to me. I'll forever miss us and cherish the shared moments. I only wish there had been more time for DnD and walks in Vicky Park.

To Fatima, for your unwavering kindness and raw sincerity. Thank you for being there for me whenever I need. I know you are there.

Pai, you were, unwillingly, the one who pushed me towards this research world, there is no way I could ever work at a bank with you as example.

Pedro, you are my rock. Thank you for being the steady presence I can hold on to. Thank you for Liverpool.

Dario, this was a journey for you too. Thank you so much for staying for the ride, for all the support and love.

Mae, you were so curious and excited when I asked you to annotate some sounds of Bees, this somehow started with you and it feels unbearable life went on without you.

For Duarte, who is the constant presence in every line of this work.

Statement of Originality

I acknowledge the ethical use of Generative Artificial Intelligence to support my editing, proofreading and/or reference list generation in this thesis. I confirm that I have kept and can provide (if requested) detailed records of my input into Generative Artificial Intelligence tools, the outputs I received, and how I used these outputs. I used the following GenAI programme(s): ChatGPT.

I accept that Queen Mary University of London has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Ines Almeida Nolasco

Date: 4th June 2025

Details of collaboration and publications: See Section. 1.5

Contents

List of Figures

List of Tables

Chapter 1

Introduction

While our curiosity about the natural world has traditionally led us to study animals at the species level, a growing body of research reveals the importance of understanding wildlife at the individual level, [Tobias and Pigot, 2019, MacKinlay and Shaw, 2023, Trappes et al., 2025]. Historically, our fascination with living things has driven us to analyse and categorise them based on heritable characteristics, drawing phylogenetic relationships across species over time to reveal evolutionary routes and provide a better understanding of current biodiversity. However, this species-centric approach, while valuable, often overlooks the rich tapestry of individual variations and behaviours within a population.

For many topics of research related to non-human animals, we tend to consider individuals as archetypes, and any variation from the species blueprint is often disregarded or treated as noise in the analysis, [Trappes et al., 2025]. This happens possibly because in regards to the questions we traditionally try to answer, these individual characteristics are not relevant. It may also be the case that until more recently we did not possess the technical capability to study animals at an individual level.

By shifting our focus to individual animals, we unlock a new dimension of understanding. This individual-based perspective allows us to explore unique behavioural patterns, social dynamics, and ecological roles that may be obscured when viewing animals solely as representatives of their species. To truly comprehend other animals, we must recognise and study their individual differences, motivations, social positions, and inherent limitations. As eloquently stated by Safina [2018], "It doesn't just matter *what* you are (...) It matters *Who* you are." This perspective invites us to consider a range of individual variations, including personalities, physical traits, learned behaviours, and life experiences. These individual differences can significantly impact an animal's interactions with its environment and conspecifics, [Réale et al., 2007]. For instance, varia-

tions in boldness might affect foraging patterns, while differences in social skills could influence mating success or group dynamics, [Carter et al., 2014] By acknowledging these individual variations, we gain a more nuanced understanding of population dynamics and social structures within species.

In human studies, the importance of individual differences is well-established, and we have witnessed the pitfalls of relying on overly broad, often male-centric definitions of *homo sapiens sapiens*, [Perez, 2019]. Similarly, in animal research, recognising and studying individual variations can provide a more accurate and comprehensive picture of species behaviour and ecology. Fortunately, technological advancements are enabling this shift towards individual-level analysis, [Dell et al., 2014, Whitford and Klimley, 2019, Nathan et al., 2022]. As our remote sensing capabilities improve, we can observe, listen to, and collect data from increasingly remote locations over extended periods. Simultaneously, our enhanced ability to process vast amounts of data allows for more efficient and detailed analysis of these diverse sources, facilitating the study of individual animals in their natural habitats.

In this work, we extend the consideration of individual characteristics to the study of animal acoustic communication and bioacoustics - the analysis of sounds that animals produce. Traditionally, species-level studies of animal communication have focused on documenting repertoires for each species, describing the acoustic characteristics of prototypical calls, and mapping the correspondence between call types and functions, [Garcia and Favaro, 2017] . However, when we apply an individual-level perspective, remembering that these sounds are produced by individuals with unique characteristics, we create space to consider variations within fixed call categories. This approach allows for novel questions and uncovers new aspects of animal behaviour and communication. For many species, we now recognize the crucial role that individual vocal characteristics play in acoustic identification of conspecifics. This capability forms the basis of some of the most critical and remarkable survival strategies across diverse species, [Carlson et al., 2020]

The evolution of technology has brought significant changes to the study of animal communication and bioacoustics. A major breakthrough has been the adoption of Machine Learning (ML) and Deep Learning (DL) techniques, [Stowell, 2022]. ML, a subset of Artificial Intelligence (AI), is designed to find patterns in data without predetermined rules. For instance, given enough examples of a cat's "meow," the machine "learns" to recognize a "meow" without explicit rules about the call's characteristics.

The widespread use of ML methodologies has automated many fundamental tasks. Call detection is a prime example; without automatic detection, researchers would spend excessive time manually annotating calls of interest,

resulting in a much smaller dataset compared to automated annotations. Call type, species, and sex classification are other tasks that can often be automated, [Stowell, 2018b]. However, the need for automatic computational systems goes beyond increased processing capacity or handling larger datasets. It's about expanding our ability to make sense of the world. Recent technologies provide novel sensing possibilities that can register even the most nuanced aspects of data.

Automatic Acoustic Identification of Individual animals (AIID) is a task that particularly benefits from these technological advancements. AIID can be defined as the process of recognising an individual based on their distinct vocalisations, [Knight et al., 2024, Terry et al., 2005]. Regardless of the type of individuality information encoded in the signal or its effectiveness, these signals are a key source of information that we can explore using ML methods to perform AIID.

In this work we want to develop methods that are able to perform this recognition in a natural context - "in the wild", a scenario which defines a set of challenges that the methods need to account for. Generally in the wild, we cannot impose hard boundaries regarding the species that we are going to record, the type of calls, or the individuals that will enter the targeted space. Ideally we want our system to work equally well across a multitude of different species, and within different conditions and time frames. These aspects direct the development of our methods towards a multi-species and open world classification framework. Beyond the application advantage that such general system might have, the multi-species and open world classification scenario poses a very challenging problem for the current machine learning algorithms and thus sets an interesting stage to further develop these methods.

The ability to identify and track individual animals in the wild opens up unprecedented opportunities for long-term studies, offering insights into life histories, population dynamics, and the intricate relationships between individuals and their environment. This thesis explores one powerful tool in this shift towards individual-level wildlife research: automatic acoustic identification of individual animals in their natural habitats.

1.1 Motivation

AIID represents a frontier in bioacoustics research with potentially important implications for animal behaviour studies and conservation efforts , [Terry et al., 2005]. While AIID has been successfully implemented in specific, controlled scenarios, developing a truly generalisable approach remains a significant challenge in the field , [Stowell et al., 2016]. Recent advancements in DL have greatly im-

proved various aspects of bioacoustics, demonstrating remarkable potential for analysing complex acoustic data, [Sethi et al., 2020]. This progress suggests that DL could be the key to achieving a breakthrough in AIID, potentially overcoming limitations that have hindered previous attempts at creating robust, widely applicable systems. The pursuit of AIID in multi-species and open world scenarios presents a unique set of challenges, [Knight et al., 2024]. These conditions provide an ideal testing ground for evaluating the limitations of state-of-the-art DL methods, particularly in terms of their capacity for generalisation and flexibility. Furthermore, generalising for multi-species means that the developed systems are not tailored for each species in particular. This, besides the application advantages of not having to develop and re-train individual systems for each particular situation, has the potential of revealing similarities in the ways very different animals encode individuality in their vocalisations. Another challenge when developing methods for AIID is the inherent lack of large annotated datasets. Considering the multi-species scenario, this issue can be even more crucial. Traditionally, if we aim to capture a larger range of characteristics we need to provide ML systems with larger training data. This aspect requires us to consider methods that can deal with very limited amounts of data, [Nolasco et al., 2023b]. By pushing the boundaries of these technologies, we not only advance the field of bioacoustics but also contribute to the broader understanding of DL applications in complex, real-world environments.

Beyond the computational aspects, the development of a system capable of reliably identifying individuals based on their vocalisations could have profound impacts on animal behaviour studies. Such a tool would enable researchers to track individuals over extended periods and across vast territories, offering unprecedented insights into social dynamics of various species. This capability is particularly valuable for studying species where traditional tracking methods may be impractical or invasive. Furthermore, the potential applications in conservation efforts are substantial. AIID could enhance population monitoring techniques, allowing for more accurate census data and better-informed conservation strategies. By advancing AIID capabilities, we aim to evolve the development of powerful tools for understanding and protecting the diverse vocal species that inhabit our world, while simultaneously pushing the boundaries of what is possible in the realm of bioacoustics and machine learning.

1.2 Aim

This thesis investigates and develops methods for Automatic Acoustic Identification of Individual animals (AIID) in multi-species scenarios. Our research focuses on four main objectives:

First, we explore learning paradigms that are less dependent on large quantities of annotated data. Given the challenges in obtaining extensive labelled datasets for certain bioacoustic tasks — of which individual identification in animals is a clear example — we investigate techniques such as few-shot learning and transfer learning. This exploration aims to develop methods that can perform adequately with limited annotated data, potentially making AIID more feasible in realistic scenarios. While this study is not centred around the AIID task, the findings here guide the framing of our task and methodologies adopted.

Focusing in AIID, we first establish a baseline for this task in a multi-species context. Given that few works have addressed the multi-species scenario before, and mainly focusing on bird species, the main step to accomplish this objective involves curating a multi-species vocalisation dataset annotated at the individual level. Our dataset includes vocalisations of both mammals and bird species and constitutes a novel and realistic scenario to further the study of acoustic signatures across species and taxonomic groups. Finally, we implement established machine learning methods on this dataset, identify their limitations, and define appropriate evaluation metrics. This baseline provides a foundation for comparing subsequent approaches.

Secondly, we select a set of promising modern DL methods, and adapt them to perform multi-species AIID predictions. Namely, we explore Hierarchical classification and Multi-task learning approaches given their potential for leveraging the extra information contained within the taxonomic organisation of the animals. By comparing their performance to the established baseline, we assess the potential benefits of these approaches.

Our third objective focus on the generalisation capabilities of the methods, particularly their ability to generalise to novel classes and therefore move towards operating in the open world scenario. We adopt Distance-based learning methods and design experiments to test how well the models handle previously unseen individuals or species, providing insights into their robustness and adaptability in more realistic conditions.

Through these objectives, we contribute to the advancement of AIID technology and provide insights into the application of machine learning techniques in bioacoustics. Our work addresses some of the practical constraints often encountered in ecological research and informs future directions in this field. Furthermore, developing such systems for AIID has a clear potential impact in animal behaviour studies and communication, biodiversity monitoring among others.

1.3 Thesis structure

Chapter 1 Introduction.

Chapter 2 Background and Relevant literature.

Chapter 3 Few-shot bioacoustics: Rethinking the big-data paradigm.

Chapter 4 A multi-species dataset for acoustic identification of individual animals.

Chapter 5 Multi-species AIID with Multi-task learning and Hierarchical classification

Chapter 6 Distance based learning of hierarchical embedding spaces for AIID.

Chapter 7 Conclusion and Final remarks.

1.4 Summary of contributions

This thesis advances the field of automatic acoustic identification of individual animals by proposing novel methodologies combined with empirical validation. Its principal contributions are summarised below:

1. **Unified multispecies AIID framework** We proposed and validated a framework capable of identifying individuals across multiple species within a single model, moving beyond the traditional species-specific systems approach.

2. **A multispecies dataset** A new dataset of bird and mammal vocalisations was compiled, labelled at taxon, species, and individual levels. Together with development split, we devised an evaluation setup comprising of three evaluation scenarios. These enable and promote development of solutions to operate in both closed set or open world scenarios.

3. **Hierarchical multi-task learning (H-MTL)** We demonstrated that jointly learning taxon, species, and individual classifiers improves accuracy and yields embeddings that better reflect biological structure than flat approaches.

4. **Analysis of distance-based learning approaches to learn meaningful representations.** In our analysis two hierarchical distance-based objectives, Rank-Based Loss (RBL) and Hierarchical Contrastive loss (HCL), were introduced and assessed as to their ability to model meaningful embedding spaces.

5. **Analysis of pretrained embeddings for AIID** An exploratory study identified OpenL3 as a strong representation backbone; this work quantifies the ability of the pretrained models to represent data at several levels of granularity while maintaining hierarchical consistency. We believe our assessment of pretrained models to be a first step towards building clear guidelines on how and when these models should be used.

1.5 Associated publications

Chapter 3, **Few-shot bioacoustics: Rethinking the big-data paradigm:**

1. Liang, J., Nolasco, I., Ghani, B., Phan, H., Benetos, E., Stowell, D. (2024, August). **Mind the domain gap: a systematic analysis on bioacoustic sound event detection.** In 2024 32nd European Signal Processing Conference (EUSIPCO) (pp. 1257-1261). IEEE. Liang et al. [2024]
2. Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., ... Stowell, D. (2023). **Learning to detect an animal sound from five examples**. Ecological informatics, 77, 102258. Nolasco et al. [2023b]
3. Nolasco, I., Singh, S., Vidana-Villa, E., Grout, E., Morford, J., Emmerson, M., ... Stowell, D. (2023). **Few-shot bioacoustic event detection at the DCASE 2023 challenge**. In Proceedings of the 8th Detection and Classifi- cation of Acoustic Scenes and Events 2023 Workshop (DCASE2023), pages 146–150, Tampere, Finland, September 2023a. Nolasco et al. [2023a]
4. Morfi, V., Nolasco, I., Lostanlen, V., Singh, S., Strandburg-Peshkin, A., Gill, L., ... Stowell, D. (2021). **Few-shot bioacoustic event detection: A new task at the dcase 2021 challenge**. In DCASE, pages 145–149, 2021b. Morfi et al. [2021]
5. Nolasco, I., Singh, S., Vidana-Villa, E., Grout, E., Morford, J., Emmerson, M., ... Stowell, D. (2022). **Few-shot bioacoustic event detection at the dcase 2022 challenge**. In Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), Nancy, France, November 2022 Nolasco et al. [2022]

My individual contributions consisted of the conceptual design of the few-shot bioacoustic event detection task, data curation and analysis, devel-

opment of the evaluation framework, evaluation of challenge submissions, and preparation of the main manuscripts.

Chapter 6, **Distance based learning of hierarchical embedding spaces for AIID.**

1. in Section. 6.2.2:
   Nolasco I, Stowell D. **Rank-based loss for learning hierarchical representations.** In International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022 May 23 (pp. 3623-3627). IEEE.[Nolasco and Stowell, 2022]
   In this work, my individual contributions included the conceptual design of the loss function, full implementation of the approach, development and execution of the evaluation framework, and preparation of the manuscript.
2. In Section. 6.2.1:
   Nolasco I, Moummad I, Stowell D, Benetos E. **Acoustic identification of individual animals with hierarchical contrastive learning**. In International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2025 Apr 6 (pp. 1-5). IEEE.[Nolasco et al., 2025]
   My individual contributions consisted of curating the dataset, contributing to the conceptual development of adopting hierarchical contrastive loss for the AIID problem, implementing the evaluation framework and conducting the experimental analysis. I was also responsible for writing the manuscript.

Chapter 2

Background and Relevant literature

Conceptually this work stands in the intersection between bioacoustic research and computer science, more particularly machine learning (ML). This chapter aims to describe the state of the art regarding Acoustic Identification of individual animals (AIID), highlighting the main challenges in the field and provide necessary support knowledge on relevant ML methods.

This chapter is organised around three central topics that frame the scientific and technical background of this thesis. First topic is an introduction to bioacoustics as a research field in Section 2.1. We outline the scope of questions it addresses and describe the key characteristics of animal sounds. in this section we also introduce the main computational approaches that have enabled automation and scalable analysis of animal sounds. Section 2.2 focuses on the acoustic identification of individual animals, AIID, the central topic of this thesis. We start by describing the biological support of vocal individuality and strategies for individual recognition. We review how acoustic signatures arise and vary across taxonomic groups and then examine relevant studies on automatic systems for AIID. To finalise, Section 2.3 defines key machine learning concepts and methods relevant to the rest of this thesis. Starting from some fundamental ML concepts that have not been captured before, we progressively build towards advanced topics in ML and deep learning that support the approaches developed in this work for AIID.

2.1 Bioacoustics and Animal sounds

Bioacoustics refers to the interdisciplinary field of research that studies sound in biological contexts. This is a broad term centred around the study of sounds produced by animals and that includes topics such as acoustic behaviour and animal communication, evolution of vocal productive systems, mechanisms for sound perception, among others [Stowell, 2018a, Gentry et al., 2020].

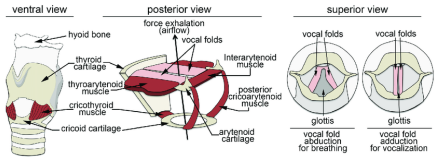
Importantly, Bioacoustics also includes the study of how sounds propagate through different environments. Understanding the physical properties of sound transmission in different habitats can be critical in order to understand aspects of the species communication. For instance, sound propagation in underwater environments considerably differs from terrestrial ones, which requires distinct analytical approaches [Ladich and Winkler, 2017]. Additionally, analysing how environmental factors affect the production and propagation of animal sounds can also provides important insights into the impacts of anthropogenic noise [Proppe and Finch, 2017].

As such, bioacoustics plays an important role in ecology, behavioural research, and conservation, offering non-invasive tools for monitoring biodiversity and ecosystems [Laiolo, 2010].

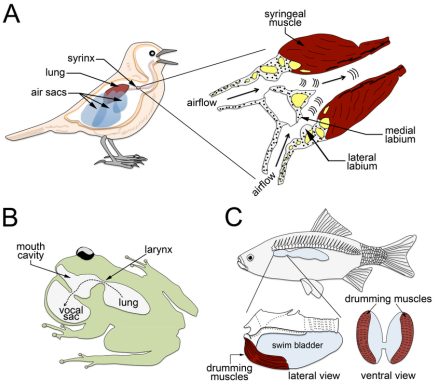
2.1.1 Sound production systems across taxa

Animals exhibit a vast diversity of vocal production systems, reflecting the wide range of anatomical structures and ecological adaptations found across taxa. This diversity equally contributes to a great variety of acoustic signals produced.

In terrestrial mammals, including humans, vocalisations are typically produced using the larynx. One established framework for understanding mammalian vocal production is the source-filter model, For instance, in terrestrial mammals vocalisations are typically produced using the larynx. A widely accepted framework for understanding mammalian vocal production is the source–filter model, originally developed for human vocalisations. This model describes vocal signals as resulting from a two step process, where the signal is first generated in the larynx (source) and further filtered while passing through the vocal tract (filter) [Taylor et al., 2016]. Figure 2.1a provides a schematic of the anatomic structures involved in the production of sounds. This model highlights how anatomical variation influences sound characteristics. For instance, the fundamental frequency is primarily determined by the size and tension of the vocal folds in the larynx and both of which vary with body size, sex, and age [Taylor and Reby, 2010]. Such factors account for the broad range of vocal characteristics observed across mammals, from infrasound calls in elephants to ultrasound



(a) Anatomy of a mammal larynx, showing ventral, posterior, and superior views of the organ.



(b) Vocal organs in [A] birds, [B] amphibians and [C] fish.

Figure 2.1: Comparative anatomy of the vocal organ in vertebrates. Reproduced from Hernandez-Miranda and Birchmeier [2018]: *Mechanisms and Neuronal Control of Vocalization in Vertebrates*, published in *Opera Medica et Physiologica* under open access.

vocalisations in rodents and bats [Ladich and Winkler, 2017].

Contrastingly different, vocal production systems in birds are described through a two-source model. The syrinx [Goller, 2022] is the organ responsible for generating vocal signals in birds. Unlike the larynx in mammals, the syrinx often contains paired structures, allowing many bird species to produce signals with multiple fundamental frequencies simultaneously [Nowicki and Marler, 1988] (see sub-figure **A** in Figure 2.1b). Also, many bird species possess accessory structures that help maintain airflow and enable further modulation of the sound. These additional adaptations contribute to the high vocal complexity observed in birds, [Ladich and Winkler, 2017].

Habitat also plays a significant role in shaping vocal production systems. For instance, semi-aquatic animals such as amphibians and cetaceans have evolved specialised structures that allow them to produce sounds without relying on a continuous external airflow. These adaptations often include air-storage organs, such as vocal sacs, that facilitate sound production even while submerged.

Despite anatomical differences, the systems described share a common prin-

ciple: sound is typically generated by forcing air through a valve-like structure, such as the larynx in mammals or the syrinx in birds, which generates acoustic signals. However, numerous other species use entirely different mechanisms for producing sound. One such method is *stridulation*, commonly observed in arthropods, where signals are generated by rubbing specific body parts together. Unlike air-based vocalisations, stridulation primarily produces vibrational signals that are transmitted through substrates rather than air [Davranoglou et al., 2023, Markl, 1965]. *Drumming* is another vibrational sound production strategy, observed across a variety of taxa from birds and mammals, to insects. In this case, repetitive physical impacts, such as pecking or hitting a surface, generate substrate-borne signals used for communication [Randall, 2010, Clark, 2016]. Fish also exhibit a sort of drumming mechanism, in which they can produce sound by contracting specialised *drumming muscles* against their swim bladder. This generates low-frequency vibrations that propagate through their body and the surrounding water [Hernandez-Miranda and Birchmeier, 2018].

2.1.2 Animal vocalisations and phylogenetic tree

Evolutionary biology determines that closely related species often share physical similarities. This principle extends beyond morphological traits to behavioural patterns, including acoustic communication. Vocalisations, being shaped by both anatomic structures of vocal production systems and behaviour, can have strong genetic components. Consequentially, species within the same taxonomic groups often share vocal characteristics that reflect the genetic similarities. Several studies demonstrate the presence of philogenetic systems in the vocalisations of species across several taxa.

Philogenetic signal is the tendency that traits from closely related species share more similarities than with randomly selected pairs. This signal can be measured through various proposed metrics which quantify how strongly trait similarities among species allign with the distance in a given phylogenetic tree [Münkemüller et al., 2012].

Several studies have shown the presence of phylogenetic signals in the acoustic properties of animals vocalisations across various taxa [Gerhardt, 1994], [Rivera et al., 2023], [McCracken and Sheldon, 1997]. In general, basic acoustic properties such as fundamental frequency and syllable structure tend to follow the phylogenetic relationships. Importantly, in Arato and Fitch [2021], the authors demonstrate that the same observation holds even in cases involving learned vocalisations, such as in songbirds.

Such findings support the hypothesis that it may be possible to organise a wide rang of animal vocalisations using a structure analogous to a taxonomic

tree.

2.1.3 Computational bioacoustics with machine learning

Computational bioacoustics refers to the application of computational methods to address research questions related to the study of animal acoustic communication and sound production. While any digital tool used to analyse bioacoustic data may fall under this term, the field has become increasingly aligned with machine learning (ML) and data-driven approaches. An important factor that led to the adoption of ML approaches to address bioacoustic questions is the growing need to process and analyse large quantities of data. The widespread deployment of affordable and automated recording devices in the field resulted in a considerable increase in data volume and complexity, which traditional manual analyses can not handle [Kershenbaum et al., 2025].

Early adoption of ML methods in this context were primarily focused on automating labour-intensive, or expensive tasks. For instance, the detection and segmentation of vocalisations of interest in long recordings, determining the presence of certain species, and species classification [Armitage and Ober, 2010, Kalan et al., 2015, Potamitis et al., 2007]. The success achieved in those early applications together with the emergence of Deep learning methods, demonstrated the potential of computational approaches to tackle questions beyond simple automation of human processes and processing of previously intractable data. In general, the tasks addressed became both more complex or nuanced such as sexing [Martin et al., 2022a], individual identification [Adi et al., 2010, Vieira et al., 2015b], acoustic monitoring of animal health and stress states [Axiu et al., 2021, Nolasco et al., 2019].

Current goals in computational bioacoustics are connected with the need to create truly generalisable systems that are less dependent on large quantities of annotated data and can operate across various conditions, species and even generalise to unseen classes [Stowell, 2022]. Within ML, there are several learning paradigms that promote these generalisation goals: Transfer Learning from large pretrained models [Ghani et al., 2023], Multi-task learning [Martin et al., 2022a], and representation learning to learn data representations that can generalise across different contexts [Moummad et al., 2024]. Initiatives such as the Few-shot bioacoustic event detection task [Nolasco et al., 2023b], the development of the first benchmarks for bioacoustics [Hagiwara et al., 2023a] and large foundational model [Robinson et al., 2024], clearly show this direction.

2.2 Acoustic Identification of Individuals (AIID)

Acoustic identification of individual animals refers to the ability to distinguish and recognise conspecifics based solely on the acoustic properties of their vocalisations [Knight et al., 2024]. The fundamental principle behind this process is that each individual must exhibit a set of vocal features that are sufficiently distinct and stable to serve as a unique identifier, these features are collectively referred to as *acoustic signatures (AS)*. In this section we look at the many forms AS take, and define the main challenges for the construction of automatic systems for AIID.

2.2.1 Acoustic Signatures

An acoustic signature can result from the effect that individual anatomic/morphological variations in the vocal system have in the signals being produced, often named as passive. As described in Gamba et al. [2017] for the case of ring-tailed lemurs, the individual variations in vocal tract morphologies between individuals account for the individual distinctiveness of their calls. AS can also be the result of an active control of the vocal system to introduce intended variations to the acoustic signal. This, in contrast, is called an active AS since it results from the active control of the muscles and structures of the vocal production system. Call-type dependent AS can often be an example of active AS, since different signatures are used depending on the function of the vocalisation. In Elie and Theunissen [2018] the authors describe how each call type of the zebra finches has its own signature features.

Due to the diversity of vocal production systems across taxa, the acoustic parameters that encode identity are not universally shared. In mammals, vocal production is typically descried through the source-filter modes Taylor and Reby [2010], and that is contrastively different from the two voice model often used to describe vocal production in birds Nowicki and Marler [1988]. Furthermore, two species with very similar vocal systems may also not encode the individual information in the same way due to requiring different strategies to perform individual recognition or identity advertisement. In Lengagne et al. [2001], the authors compare the individual recognition process of King penguins and their close species Emperor and Adelie penguins. Contrary to those, in King penguins all the syllables of a call contain the identity information. While penguins in others species need to listen to several syllables before the identity of the caller is ascertained. The authors hypothesise that this fundamental difference in identity advertisement may be due to the different society organisation and habitat of each species. The King Penguins live in colonies with millions of individuals in what becomes very noisy conditions. The encoding of the identity in

the vocalisations seems to have evolved accordingly, in a way that optimises individual recognition in that environment. Another study Mathevon et al. [2003], comparing two species of gulls that present different nesting patterns, describes similar findings where the different behaviour and conditions determines very different AS. In this case the species where the requirement for identification of the parents is higher developed an AS based on two fundamental frequencies which gives the signal larger capacity to transmit the identity information. Also in Charrier et al. [2009] the authors study how the environment conditions and societal organisation of sea lions might impact their acoustic signatures. Colonial living species as the sea lions have typically many individuals using the same communication channels and this imposes a strong selection pressure on the development of their communication systems. Variation of AS within the same species is also possible. As the study in Smith-Vidaurre et al. [2021] suggests, the size of the social group or population where the monk parakeets live determines the complexity of their AS.

Numerous studies have aimed to identify the specific acoustic features that convey identity across taxa. For instance, in wolves, the fundamental frequency of howls carries individual-specific information [Root-Gutteridge et al., 2014]. In many songbirds such as tree pipits, identity is encoded through song syntax and repertoire content[Petrusková et al., 2016]. Domestic sheep use a combination of mean frequency, spectral distribution, and timbre for mother-offspring recognition [Searby and Jouventin, 2003]. Given the wide variety of behaviours and anatomical differences it becomes challenging to produce a unified description of the features used to encode identity.

The question of stability of the AS over time is also important to consider. Some species continue to develop their repertoire of calls over the course of their lives which in many cases leads to AS that either change or are circumstantial to a situation or time frame Elie and Theunissen [2018]. For other species, AS remain constant for long periods of time, as is the case of northern fur seals. These have been shown to have the capacity for long term recognition between mothers and previous pups Insley [2000].

As a note, for the purpose of automatic acoustic identification of individuals, we do not necessarily need to work with species that perform individual recognition. Indeed presenting individualised acoustic features that can be used for individual identification does not imply that the animals perform recognition within their own natural behaviour. However, it is very likely that the development of acoustic signatures and the development of strategies for recognition of conspecifics to have happened through an intertwined evolutionary process, as described for the sea lions in Charrier et al. [2009]. Understanding the latter can thus help understand the first.

2.2.2 Measuring Vocal Individuality

The work by Linhart et al. [2019] presents a different approach to the study of acoustic signatures from an information theory side. Here the authors focus on testing which is the best metric to measure the individual information encoded in the signal, and thus propose a clear way of evaluating the strength of an acoustic signature. The main idea behind these metrics is that an AS must maximise the potential for encoding identity in the signal.

The Beecher information (H_S) by Beecher [1989], is such a metric, it measures the information capacity of a set of calls, indicating how individually distinctive they are. H_S is based on the ratio of between-individual variation to within-individual variation in a particular trait of the signal. The greater the between-individual variation in relation to the within-individual variation, the greater the amount of information present and the easier it is to distinguish between individuals.

2.2.3 Automatic approaches to AIID

The diversity of acoustic signature characteristics, their non stability, and the lack of a unified model for encoding identity in the vocalisations across different taxonomic groups present significant challenges for developing automatic systems capable of AIID. These challenges remain key obstacles in building general-purpose, scalable AIID systems and have led most previous work to focus on narrow, species-specific contexts.

Early efforts in AIID employed statistical models based on engineered acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), formant frequencies, or energy contours. Common techniques include Gaussian Mixture Models (GMMs) [Cheng et al., 2010], Discriminant Linear Analysis (DLA) [Sadhukhan et al., 2021, Yin and McCowan, 2004], and Hidden Markov Models (HMMs) [Wijers et al., 2021, Vieira et al., 2015a]. These approaches typically required careful feature engineering and were often limited in their ability to generalize across individuals or species. The work of Stowell et al. [2019] marked one of the first explicit attempts to address AIID across multiple bird species and varying recording conditions, while also highlighting the issue of confounders — spurious correlations between vocal identity and factors such as background noise, location, or recording equipment.

More recent developments have adopted deep learning frameworks to overcome many of these limitations. Martin et al. [2022a] introduced a multi-task learning approach based on convolutional neural networks (CNNs), simultaneously predicting the animal's sex, performing call detection and and individual classification.

A growing body of work also explores AIID in open-world or open-set scenarios, where individuals unseen during training may occur during testing [Knight et al., 2024]. One example is the work by Ntalampiras and Potamitis [2021], who employed variational autoencoders (VAEs) to model the latent space of known individuals and identify novel individuals as outliers based on reconstruction error. Beyond supervised classification, unsupervised and clustering-based approaches have also been adopted in this context. Bedoya and Molles [2021] applied deep clustering algorithms to directly discover individual vocalizations without requiring labelled data, enabling operation in the presence of unknown classes. Similarly Guerrero Muriel et al. [2023] employed unsupervised deep clustering methods to identify individual birds from large-scale passive acoustic recordings. An important contribution towards open-world AIID was presented by Huang et al. [2024], who developed an individual recognition system capable of handling both known and novel individuals in five bird species. Their system leverages pretrained embeddings derived from species-level classification models, which improves individual separability and allows for better generalization to unseen individuals.

As computational bioacoustics enters the era of foundation models and general-purpose systems, there is an increasing need to evaluate these models across diverse bioacoustic tasks. AIID is particularly valuable in this context, as it reflects both the fine-grained discrimination and generalisation capabilities of such systems. Consequently, AIID has been adopted as an evaluation task in recent initiatives such as Hamer et al. [2023] and Miron et al. [2025]. More specifically, the *BEANS* benchmark dataset introduced by Hagiwara et al. [2023b] includes data from dogs and bats, each with individually labelled recordings. Despite the growing prevalence of systems capable of performing AIID in particular contexts, there remains a lack of structured approach and methodological framework for AIID that address key challenges such as generalisation across multiple species and open-set recognition.

2.2.4 Individual identification from other modalities

individual recognition of conspecifics is rarely based on a single sensory cue; instead, it typically relies on a combination of auditory, visual, and olfactory information that animals can associate with unique individuals [Thom and Hurst, 2004, Yorzinski, 2017].

Automatic approaches to individual identification are still far from integrating these multiple modalities. However, unimodal systems have been successfully developed. In the visual domain, extensive research has focused on using image-based features such as body shape, fur or skin patterns, and facial mark-

ings to distinguish individuals [Vidal et al., 2021, Norouzzadeh et al., 2018]. This task is commonly referred to as animal re-identification (re-id), drawing parallels with person re-identification in computer vision. More recent work, such as Čermák et al. [2024], illustrates a shift in the re-id field toward cross-species generalisation. Their MegaDescriptor system learns a shared visual embedding trained to discriminate between individuals across multiple species.

Environmental DNA (eDNA) represents another modality that has demonstrated potential for identifying individuals and for characterising and monitoring populations within ecosystems, [Piggott and Taylor, 2003, Broadhurst et al., 2025]. In particular scenarios, individual animals can also be distinguished by their characteristic movement patterns, such as gait or flight trajectories, [Teshima et al., 2025].

2.3 Machine Learning Methods and Advanced Topics

In this section we aim to provide the necessary knowledge base for the machine learning approach, and describe the learning paradigms and methodological tools used through out this thesis.

Machine learning (ML) refers to a set of computational techniques that allow systems to learn patterns directly from data, rather than relying on explicit rules or instructions. As described by Goodfellow et al. [2016], machine learning enables computers to acquire knowledge from experience, observed through patterns present in the training data. By repeatedly encountering similar examples, algorithms learn statistical relationships, enabling predictions on new, previously unseen data examples.

2.3.1 The supervised learning pipeline

There are various learning paradigms that realise this process of machines learning from data. **Supervised learning** is such a paradigm characterized by the presence of labelled data, where each training example is associated with the desired output, the goal is to learn a mapping from input features to output labels [Bishop and Nasrabadi, 2006].

Through this process different types of tasks can be defined. In the audio domain we define **classification** or audio tagging as the task of predicting categorical labels that describe the audio segment in the input. **Sound Event detection** is another type of task, where the goal is to find the onset and offset of certain audio events [Virtanen et al., 2018].

A standard supervised learning pipeline typically comprises several stages:

1. **Feature Extraction:** Computation of relevant characteristics from input data to solve the task at hand; The feature extraction step is a crucial one, it can be described as where the data is transformed and its dimensionality reduced in order to fit the training algorithm used. Features need to both simplify the raw signals in the input and preserve the key information for the target task. They can also be a way to highlight the characteristics of the signal which we want the model to focus on. In short, this is where we define how the signals are represented. In traditional ML application, features used tended to be more high level summarisations of the characteristics of the signal, and often highly engineered. As the field evolves and typically for deep learning, representations tend to be more low level, raw and with less human bias introduced. This is both because the computational capacity increased but also because the models have an increased capacity for learning at various levels of detail and can cope with their levels of information. Typically these representations take the form of **Spectrograms** – frequency-time representations of the wave signal. A frequent used variation is **Melspectrograms** which modifies the frequency scale to align better to how humans perceive sound [Virtanen et al., 2018]. An important development regarding feature extraction is using the internal representation of another model – **embeddings** – as the initial input for further learning.

   This approach is tightly also connected with the concept of **representation learning**. Representation learning is a process by which a machine learning model automatically discovers and transforms input data into informative features or representations that are useful for the learning task [Bengio et al., 2013].

2. **Model training:** The learning algorithm is trained on the training set, hyper parameters are adjusted depending on performance on the validation set; In this work, step two is based mainly on the training of **Neural Networks (NNs)**. Neural networks consist of interconnected layers of artificial neurons that can learn hierarchical data representations directly from input features [LeCun et al., 2015].

   Neural networks are trained in iterations of forward and back propagation, and use optimization algorithms, such as **stochastic gradient descent** (SGD), that minimize a loss function. The loss function defines what is the outcome we want to achieve, in classification tasks a common loss function is the **cross entropy loss** which penalises the system when predictions and ground truth labels do not match [Goodfellow et al., 2016].

   For binary classification, the cross-entropy loss is given by:

$$\mathcal{L}\text{BCE} = -\sum i = 1^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{2.1}$$

where y_i is the true binary label and $\hat{y}_i$ is the predicted probability of the positive class.

The cross entropy loss can be extended to multi-class classification tasks, taking the form:

$$\mathcal{L}\text{CE} = -\sum i = 1^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}), \tag{2.2}$$

where N is the number of examples, C is the number of classes, $y_{i,c}$ is the true label indicator (1 if the i^{th} sample belongs to class c, otherwise 0), and $\hat{y}_{i,c}$ is the predicted probability of the i^{th} example belonging to class c. Minimizing these loss functions encourages the model to produce outputs closely aligned with the true class distributions, and during the training process, the optimisation algorithm finds the set of parameters (weights) that best minimise the loss function.

3. **Evaluation:** The trained model is applied on the test set and evaluation metrics measure the quality of the predictions performed.

   In a successfully training process, as the parameters are adjusted, this loss will tends towards zero. Monitoring the loss is a first step to assess the success of the training process. Another way to assess the effectiveness of the training is to evaluate the actual predictions that the trained model produces by comparing them to the ground truth labels. For this, we use **evaluation metrics** such as accuracy, precision, recall, and F-score among others.

   Accuracy is the proportion of correct predictions over the total number of predictions made:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}. \tag{2.3}$$

   Precision measures the proportion of true positive predictions among all samples predicted as positive and reflects how many of the positive predictions are actually correct:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}. \tag{2.4}$$

   Recall quantifies the proportion of true positives correctly identified out

of all actual positives, it reflects the model's ability to capture all positive instances:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}. \tag{2.5}$$

F-score is the harmonic mean of precision and recall, providing a balanced metric that accounts for both:

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.6}$$

Balanced accuracy when dealing with unbalanced datasets, traditional averaged accuracy metrics may not provide a complete picture of model performance. Instead the use of balanced accuracy metric is preferred which is based on the average of recall per class.

$$\begin{aligned} \text{Balanced Accuracy} &= \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \\ \text{where} \quad & C \text{ is the number of classes,} \\ & \text{TP}_i \text{ is the number of true positives for class } i, \\ & \text{FN}_i \text{ is the number of false negatives for class } i. \end{aligned} \tag{2.7}$$

The training loss and evaluation results on the test set should, up to a point, follow similar trends. However, it is important to strike a balance, as we aim to avoid the model overfitting to the training data. **Overfitting** occurs when the model becomes overly specialized in the training set, effectively memorizing its detailed characteristics rather than learning patterns that generalize to unseen data from the same classes. As a result, while the model may achieve very low training loss, its performance on new data degrades.

To mitigate overfitting, one commonly used technique is **early stopping**. In this approach, after each training epoch, the model's performance is evaluated on a separate validation set. If the validation performance starts to deteriorate which indicates that the model is no longer improving performance on unseen examples, training is stopped. This allows to find the correct balance between training and minimising the loss function and generalisation ability to new examples.

These steps summarise the supervised learning process with neural networks, while other configurations and learning objectives exist, pipelines typically some

version of the 3 stages described above. In the next sections we present such alternative configurations and learning paradigms that are relevant to the present work.

2.3.2 Multi-Task Learning (MTL)

Multi-task learning (MTL) is a machine learning paradigm where multiple related tasks are learned simultaneously rather than in isolation. The primary objective is to develop a model that generalizes better across tasks by leveraging shared knowledge. This means that the learning goal is a combination of several objectives and the model learns to prefer solutions that in general satisfy all objectives.

This unified training goal is captured in the **MTL loss function** which consists on a weighted sum of task-specific losses:

$$L_{MTL} = \lambda_1 L_{task1} + \lambda_2 L_{task2} + \lambda_3 L_{task3} \tag{2.8}$$

where each L_{taski} represents the loss for a given task (cross entropy or other), weighted by the task-specific coefficient λ_i. The choice of λ_i values allows controlling the desired impact of each individual task on the total loss. In other words, by increasing λ_i for one of the tasks we can focus more on that task than others.

A key advantage of MTL is that it encourages the network to learn latent representations that are useful across tasks. This has the potential to improve generalization, particularly in low-data scenarios where related tasks can help compensate for limited individual datasets. The rationale behind it is that what the first layers of a neural network represent is general enough to be shared among different but related tasks. More importantly is the observation that if this shared representation is learnt together, *i.e,* it learns to optimise for the different tasks at the same time, then the models will generalise better at each task than if instead they had been learnt in isolation.

Various configurations of neural networks exist that define how strict is the transference of knowledge across tasks and joint optimisation. **Hard parameter sharing** [Ruder, 2017] is such a configuration where each task contributes to optimise a common branch (illustrated in Fig. 2.2. In this setup, the common branch, where the shared representation across tasks emerges, is followed by task-specific branches that perform classification for the individual tasks.

During training, the total loss is back-propagated and used to update the network. Shared layers are updated from the gradient steps of the total loss, whereas task-specific layers are updated only from the gradient of the respective task loss This structured learning is responsible for making the shared layers

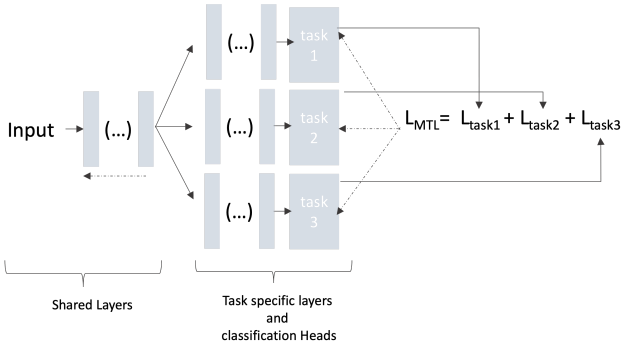


Figure 2.2: Common architecture for a Multi-task learning network with Hard parameter sharing. Each task specific branch contributes to the total Loss (L_{MTL}). Dashed lines represent the back-propagation flow of the loss. shared layers are updated from the gradient steps of the total loss, whereas branch-specific layers are updated from the gradient of the respective task loss.

capture a more general representation while the task-specific layers can specialise and fine tune to their specific objectives.

This configuration contrasts with **soft parameter sharing** [Ruder, 2017], where each task maintains a separate network, and similarity across tasks is encouraged through regularization penalties that keep the parameters of corresponding layers close. In soft sharing, no explicit shared representation is enforced at the early stages of the network.

When designing MTL approaches, two key considerations should be addressed:

First, it is necessary to determine which tasks should be jointly learnt. How to choose the set of tasks that will promote mutual benefit and effective knowledge transfer is an ongoing research problem. In general tasks that are related may reinforce shared representations, while unrelated or conflicting tasks can hinder learning. Techniques such as task clustering have been proposed to guide this selection Zhang and Yang [2017].

Second, is how the input data is fed to the MTL network. The most common scenario assumes the existence of a single dataset where each data sample is labelled for all tasks simultaneously, which means that every example is associated with a label for every task. However, It is still possible to design a MTL approach if there is not a single fully labelled dataset. An alternative is to train with separate datasets for each task, where each dataset contains labels specific to its corresponding task. In this case, the training process typically alternates between tasks, updating the network parameters based on the available data for each task in turn. This alternating training strategy introduces additional sensitivity to dataset imbalances, as tasks with larger datasets may dominate

the optimization process, potentially biasing the shared representation.

2.3.3 Distance-Based and metric learning

In standard supervised learning, the model learns internal feature representations as a by-product of optimizing for the classification objective. Through the training process, the network transforms input signals into progressively higher-level features that facilitate class separation. Specifically, for classification tasks, the model adjusts its internal representation to increase separability between classes, improving its ability to assign correct class labels. While this process can produce embedding spaces that reflect some underlying semantic structure this organization is not explicitly controlled. The embedding space emerges indirectly from the classification objective, and there is no guarantee that the learned structure will generalize well to different but related tasks, or that it will capture meaningful relationships beyond the classes seen during training.

By contrast, **metric learning** directly formulates the semantic organization of the embedding space as the objective of training [Bellet et al., 2013]. Rather than optimizing class probabilities, metric learning aims to arrange samples in the feature space such that semantically similar examples are positioned close together, while dissimilar examples are pushed farther apart. This is typically achieved by minimizing loss functions that operate on comparisons between pairs or triplets of samples, such as contrastive loss or triplet loss.

Closely related to metric learning is the concept of **distance-based classification**, which focuses primarily on the prediction stage. Distance based methods, such as k-nearest Neighbors or prototypical networks [Snell et al., 2017], classify new samples by computing distances to labelled reference points or class prototypes in the embedding space. In this thesis, we use the term Distance based learning (DBL) to encompass both the learning of embedding spaces and the subsequent prediction based on distance comparisons.

Contrastive Learning

Contrastive learning aims to map semantically similar data points closer together in the embedding space while pushing dissimilar points further apart. At its core, it is a learning paradigm based on direct comparisons between examples Instead of optimizing for class probabilities (as in standard classification), contrastive learning optimizes for the relative positioning of samples in a learned feature (embedding) space [Le-Khac et al., 2020].

The degree of similarity between samples in the embedding space is measured using a similarity function, often related to the distance between their

embeddings, such as Euclidean distance or cosine similarity. The formulation of contrastive learning into a training objective (or loss function) depends on how semantic similarity is defined, whether similarity labels are derived from ground truth class labels (supervised) or generated without labels (self-supervised). In all cases, positive pairs are formed by semantically similar examples, while negative pairs consist of dissimilar examples.

The supervised contrastive loss [Khosla et al., 2020] operates by selecting, for each anchor sample i, all other samples with the same class label in the batch as positives. The loss encourages the model to pull together embeddings of all positive pairs while contrasting them against all other samples in the batch.

$$L_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{n \in N(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_n / \tau)} \tag{2.9}$$

Where I denotes the set of indices for all samples in the batch, and $\mathbf{z}_i$ represents the embedding vector of sample i. The parameter τ is the temperature scaling factor. The set of positives for anchor i is given by $P(i) \subset I \setminus \{i\}$, containing all samples in the batch that share the same class label as i. The set $N(i) = I \setminus \{i\}$ includes all other samples in the batch except the anchor itself. The similarity between two embeddings is computed as $\mathbf{z}_i \cdot \mathbf{z}_p$, corresponding to the dot product between the anchor and a positive sample (equivalent to cosine similarity when the embeddings are normalized).

More recently, self-supervised contrastive learning has been proposed, where similarity is defined without class labels [Chen et al., 2020b]. In this setting, positive pairs are formed by generating two augmented views of the same input sample, while all other samples in the batch serve as negatives.

Triplet Loss

Triplet loss shares the same fundamental principle as contrastive loss, in that it is based on learning from similarity comparisons across examples. However, instead of relying on pairwise comparisons, triplet loss operates on triplets of samples, where the relative distance between positive and negative examples is directly controlled through a margin parameter.

Each triplet consists of an an *anchor* sample x_a, a *positive* sample x_p (belonging to the same class as the anchor), and a *negative* sample x_n (belonging to a different class). The objective is to ensure that the distance between the anchor and positive is smaller than the distance between the anchor and negative by at least a margin α. Formally, the triplet loss is defined as:

$$\mathcal{L}\text{Triplet} = \sum i = 1^N \max\left(0, ; |f(x_a^i) - f(x_p^i)|_2^2 - |f(x_a^i) - f(x_n^i)|_2^2 + \alpha\right), \tag{2.10}$$

Where $f(\cdot)$ is the embedding function, $|\cdot|2$ denotes the Euclidean distance, α is the margin parameter, and $\max(0, \cdot)$. Minimizing this loss encourages the model to embed positive pairs closer together while pushing negative pairs further apart, enforcing that positive pairs are separated from negative pairs by at least the margin α. As a result, the learned embedding space directly reflects semantic similarity through its distance relationships

2.3.4 Tools for evaluating embedding spaces

Using pretrained models as embedding generators, or explicitly learning embedding spaces that capture some form of geometric and semantic structure, has become an increasingly common practice across ML applications and domains. However, there remains no standardized framework for evaluating and selecting embedding spaces, nor established strategies to determine which embedding structures are best suited for a given downstream task. In the audio domain a few initiatives have attempted to address this issue, such as Yuan et al. [2023], Turian et al. [2022], however evaluation strategies remain to be on a case-by-case basis, where models are often selected based on performance on specific downstream tasks or even on the practicality of extracting embeddings from them.

In this section, we describe a set of tools and metrics that are used in this work for assessing and comparing embedding spaces across models.

K-nearest neighbours (kNN)

kNN is a distance based classification algorithm that assigns a class label to a new data point by identifying the K closest training samples in the embedding space and selecting the most frequent class among them as the label to assign. The closeness between points is typically measured using distance metrics such as Euclidean distance , cosine similarity or others.

Formally, given an input sample x, the algorithm computes the distances between x and all training samples, selects the set of K nearest neighbours, and predicts the class label by majority vote among these neighbours. The parameter K influences the smoothness and robustness of the classification outcomes. [Bishop and Nasrabadi, 2006]

As an evaluation tool, kNN can easily demonstrate if the embedding space provides enough separability across classes that promotes straightforward clas-

sification. It also offers some flexibility in regards to the choice of k or the reference set, allowing for more nuanced insights into the structure of the embedding space

Silhouette scores

Focusing on the structure or geometry of the embedding space, cluster quality metrics can provide quantitative measures of how well represented the classes are in the space. Silhouette scores are a commonly used metric, which captures both the compactness of individual clusters and the separation between clusters in the embedding space [Shahapure and Nicholas, 2020].

Formally, for each sample i, the silhouette score $s(i)$ is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i),\, b(i)\right)}, \tag{2.11}$$

where $a(i)$ denotes the mean intra-cluster distance, i.e., the average distance between sample i and all other samples assigned to the same cluster, and $b(i)$ denotes the mean nearest-cluster distance, i.e., the average distance between sample i and all samples assigned to the closest neighbouring cluster. The silhouette scores ranges between -1 and 1, where higher values indicate better cluster separation.

The average silhouette score across all samples provides a useful global measure of the overall cluster structure, offering insight into the separability of classes in the embedding space.

Visualisation of embedding spaces

While silhouette scores and kNN classification performance provide concrete quantitative measures of the structure and suitability of the embedding spaces, these metrics can sometimes be challenging to interpret in isolation. In this context, visual inspection of the embedding space offers a valuable complementary tool for assessing the quality of the embedding representations. By visualizing how samples are organized in the embedding space, it becomes easier to interpret spatial distribution of the classes, class separability and overlap.

Dimensionality reduction algorithms are used to project the high-dimensional embedding space into two or three dimensions, while preserving, as much as possible, the relational structure between samples. In this work we make use of two methods:

- Principal Component Analysis (PCA) is a classical linear dimensionality reduction technique that projects high-dimensional data into a lower-dimensional space by identifying orthogonal directions (principal compo-

nents) along which the data exhibits maximal variance. The first principal component captures the largest amount of variance, with each subsequent component capturing the next highest variance under the constraint of orthogonality. PCA preserves global variance structure but may struggle to capture complex nonlinear relationships between data points [Abdi and Williams, 2010].

- Uniform Manifold Approximation and Projection (UMAP) is a nonlinear dimensionality reduction algorithm that seeks to preserve both local and global topological structure of the data manifold. UMAP constructs a weighted graph representing local relationships in high-dimensional space and then optimizes a low-dimensional representation that minimizes divergence between the high- and low-dimensional structures. This makes UMAP particularly effective for visualizing complex embeddings with underlying manifold structures that may not be well-captured by linear projections [McInnes et al., 2018].

2.3.5 Hierarchical Classification

In most classification tasks, target classes are treated as independent and are typically assumed to represent concepts at the same semantic level. Even in multi-label classification setups, where each example can be associated with multiple labels, the relationships between classes are usually not incorporated into the learning process. However, in many real-world domains, this flat organisation of classes is an oversimplification. Hierarchical taxonomies, where concepts are organised in a tree-like structure from broader categories to increasingly specific and narrower subcategories, are common [Silla and Freitas, 2011].

In machine learning, methods that leverage these hierarchical relationships among classes are referred to as hierarchical classification. Such methods are attractive because they can improve performance on more challenging, fine-grained classification tasks by first solving coarser, higher-level decisions. In essence, hierarchical classification approaches improve fine-level predictions by either narrowing the set of possible classes based on earlier parent-level predictions or by weighting the predictions of child classes according to the confidence in their corresponding parent class, [Pham et al., 2020].

Classical hierarchical classification approaches [Silla and Freitas, 2011], can be categorized according to several key aspects:

Training strategy: whether a single global model is trained to cover the entire hierarchy or multiple specialized "local" models are trained for different parts of the hierarchy (i.e., *Global* vs. *Local* approaches).

Prediction sequence: depending on the training strategy, predictions can be made in a sequential **Top-Down** way, or produced jointly through global decision models that consider the entire hierarchy at once.

Label completeness: determines whether predictions are required to reach leaf nodes (i.e., full-depth classification) or whether predictions may stop at internal nodes, resulting in partial classification.

Hierarchical structure: specifies the type of hierarchy the model is designed to handle, ranging from strictly tree-structured hierarchies to more general Directed Acyclic Graphs (DAGs) that allow multiple parent nodes per class.

More current implementations of hierarchical classification mainly fit into the two paradigms introduce above:

Multi-task learning, in which we have one task per level of the hierarchy and following the multi-task learning framework , the network learns to optimise for the multiple tasks together. for example, Xu et al. [2016] targets acoustic scene classification on a 2 level tree label structure. besides the combined loss from the MTL setup, this implementation proposes a strategy for pretraining the networks on a single level of the label tree in order to improve the training and performance on other levels. In Cramer et al. [2020], the authors address bird species classification from flight calls. Here the data is organised in a 3 level taxonomy (animal order, family, species). A novel network architecture is proposed -*Taxonet*- where layers are partioned accordingly to the hierarchical strucure of the labels. and nodes are activated depending on the parent partion having been activated or not in the previous layer.

Distance based Learning, the goal is to learn embeddings that explicitly convey the hierarchical structure of the problem. in [Elizalde et al., 2018], the authors explore the use of Siamese networks and manually define a target distance between pairs of items in the generated embeddings depending on the position of the input examples on the label tree. Hierarchical information is also integrated through the network architecture by multiplying an incidence matrix with the output layer predicting the leaf-level of classes which generated the output predictions for the higher level classes. both works by Garcia et al. [2021] and Liang et al. [2022], focus on audio classification tasks by implementing hierarchical prototypical networks, here the prototypes of fine level classes are aggregated into broader prototypes that represent the classes at higher levels. Finally,

Jati et al. [2019] employs a quadruplet loss (generalisation of triplet loss) approach to sound event detection on a dataset organised in a two level hierarchical tree. The core of their proposed method is to build quadruplets that contain examples of all the possible hierarchical relationships of the label tree. *I.e.,* anchor and positive are examples from the same leaf-label, negative are examples from the same coarse level and different leaf-label, or from a different coarse level. The authors report improved classification results at both levels.

2.3.6 Few-shot learning

Supervised learning is typically supported by extensive datasets that capture the wide variability of real-world contexts, providing the model with enough labelled examples to identify patterns and generalize from training to unseen data.

Few-shot learning (FSL) emerges from the move in ML research towards creating models that are less dependent on large and all encompassing datasets and consequentially can operate under limited supervision. FSL problems are defined as learning to perform a task given very few labelled examples, typically one or five instances per target class [Wang et al., 2020a]. In contrast to conventional supervised learning, FSL focuses on achieving generalization from minimal data and it requires for generalization across tasks, *i.e.,* the ability to make use of prior experience from well-supported tasks to solve novel tasks.

Main approaches to FSL can be split in two: (1) **Transfer learning** or **Representation learning** and (2) **Meta-learning**.

In transfer learning and representation learning approaches, the underlying idea is that a foundational representation of the domain can be learned from large datasets and then transferred or fine-tuned to novel tasks with limited data. This is commonly achieved by training models to generate embeddings that capture meaningful characteristics of the data. These embeddings, either learned via supervised, self-supervised, or contrastive methods, serve as general-purpose representations that can be extended to novel classes with minimal adaptation. Recent works also make use of **transductive inference**, which defines an adaptation mechanism that integrates the unlabelled elements of the data in the process of learning an embedding space, allowing for better representation of the targeted classes even is their characteristics are not completely captured within the few shots available. For example, both Boudiaf et al. [2021], You et al. [2023] apply transductive inference techniques to better exploit unlabelled data from the target domain leading the models models to capture richer domain representations.

In contrast, meta-learning approaches attempt to solve the few-shot learning problem by explicitly learning how to learn. Inspired by human learning, meta-learning frameworks train models across a diverse set of tasks, enabling the model to capture the underlying learning process that links inputs to outputs, rather than memorizing fixed associations. As illustrated by the example in Wang et al. [2020a], a child who learns to add can rapidly extend this knowledge to multiplication by recognizing its relation to repeated addition. by observing a few examples (tasks), the dhild learns process of decomposing multiplication into a series of additions instead of memorising the association between input and outcome. e.g the association is not that 2*3 = 6 but that 2*3 = 3+3.

Similarly, meta-learning methods aim to expose models to a sufficient diversity of tasks such that they can acquire the underlying learning mechanism necessary for rapid generalization. **Episodic training**[Li et al., 2019], is a central process in Meta-learning through which models acquire this skill. Unlike conventional supervised learning, where each batch contains randomly sampled examples across classes, episodic learning structures training into a sequence of simulated few-shot tasks. Each episode is designed to mimic the actual few-shot scenario the model will encounter at test time, improving the model's ability to generalize under limited data regimes.

Formally, episodic learning typically follows an **N-way K-shot** formulation: N-way refers to the number of classes included in each episode and K-shot specifies the number of labelled examples provided per class in the support set.

For example, a 5-way 1-shot task presents the model with 5 classes, each containing 1 labelled example (the support set), followed by a small set of unlabelled query samples drawn from the same classes. The model is trained to classify the query samples (unlabelled) based on the information in the support set. This episodic structure makes the model repeatedly adapt to new tasks and consequentially improve its ability to quickly generalise to new tasks.

This training paradigm has been widely adopted in few-shot learning methods [Finn et al., 2017]. For instance, in Snell et al. [2017], the authors train a prototypical network through an episodic training process, where each episode requires the model to compute class prototypes from the support set and classify query samples based on their distance to these prototypes in the embedding space.

Prototypical networks

Prototypical networks are designed specifically as a few-shot learning algorithm, where the focus is on learning an embedding space where classification can be performed through simple distance comparisons between a query sample and

the prototypes. prototypes are learnt based on the few-shots provided and are intended to represent each class.

A *prototype* is a coordinate in some vector representation, which is calculated as the centroid of the coordinates for each of the shots. The training data consist of a *support set* S consisting of k labelled samples from each class, with the remaining samples comprising the *query set* Q. Prototypical networks compute a class prototype c_n through an embedding function $f_\phi : \mathbb{R}^D \rightarrow \mathbb{R}^M$ with learnable parameters ϕ. The prototype for class n is computed as the mean of the embedded support points belonging to that class:

$$c_n = \frac{1}{k} \sum_{(x_i) \in S_n} f_\phi(x_i) \tag{2.12}$$

where S_n represents the subset of S from class n. Then, for each sample x_q from the query set, a distance function is used to calculate the Euclidean distance of x_q from each prototype, following which a softmax function over the distances produces a distribution over the classes [Snell et al., 2017].

2.3.7 Open world Scenario

Tasks in machine learning are typically framed under the assumption of a well-defined and fixed taxonomy; that is, they are often approached using a closed set formulation, where the categories encountered during application or testing are assumed to match those seen during training. In contrast, open set scenarios acknowledge that previously unseen classes may appear at test time, classes that the model has never been exposed to during training. Models that are not explicitly designed to handle such novel classes are more prone to errors, including misclassifications and overconfident predictions, which limits their robustness and applicability in real-world deployments.

The first requirement for operating under these unconstrained conditions is to detect whether a given input belongs to any of the previously learned classes or constitutes a novel class. This is the central problem addressed by the class of methods known as Open Set Recognition (OSR) [Scheirer et al., 2013, Mahdavi and Carvalho, 2021]. OSR methods typically extend standard classification by introducing mechanisms that estimate the model's uncertainty or distance from known class boundaries, enabling detection of unknown classes while maintaining classification of known ones.

Several strategies have been proposed to implement OSR: Distance-based and metric learning approaches utilize embedding spaces to measure similarity to known classes and reject samples that fall beyond a threshold [Bendale and Boult, 2016]. Generative models, attempt to simulate variations of unknown

class examples to improve detection capabilities [Perera et al., 2020]. Other approaches employ Extreme Value Theory (EVT) to statistically model the tails of known class distributions, enabling robust rejection of samples falling into extreme regions of the feature space [Rudd et al., 2017]. In short, OSR methods leverage the ability to distinguish and isolate occurrences of novel classes in order to better separate and consequently improve classification of known classes.

However, going beyond the simple identification or rejection of novel classes, an important objective is to actively learn and incorporate these new classes into the system's classification vocabulary. This broader paradigm is referred to as Open World Recognition (OWR) [Bendale and Boult, 2015], and intersects with related areas such as Continual Learning [Wang et al., 2024] and Novel Class Discovery (NCD) [Zhong et al., 2021]. The underlying approach for OWR relies on the assumption that a semantic meaningful representation of the known classes generalises to novel classes as well, allowing novel classes to be discovered and incorporated through clustering, proximity, or similarity reasoning in a learned feature space.

2.3.8 Research directions and gaps

Most of the deep learning frameworks described in this chapter are not older than a decade [Stowell, 2022], and the majority of them have been adapted from developments in computer vision or speech processing. More recently, research in computational bioacoustics has gained autonomy and progressed substantially. However, the field is only now entering the era of foundation models and large-scale representation learning for bioacoustics [Schwinger et al., 2025, Robinson et al., 2024]. The emerging direction is towards the development of general-purpose models that can be adapted to multiple bioacoustic tasks and generalise across species, recording conditions, and acoustic domains.

There remains, however, a clear need for architectures capable of attending to very short-term acoustic events while also preserving the long-term temporal context of recording sessions. Transformers have been increasingly adopted in the audio and bioacoustics domains with demonstrated success [Kather et al., 2025]. These networks learn relationships between elements in a sequence using self-attention, which makes them particularly effective at modelling long-range dependencies and contextual structure. Nevertheless, their computational cost remains high.

Another major research direction concerns reducing dependence on heavily annotated datasets. Completely unsupervised learning applied to bioacoustic data has not yet achieved consistent success. This is largely due to the high variability of acoustic characteristics arising not only from individual-level dif-

ferences but also from the use of different recording devices and environmental noise conditions. As a result, unsupervised methods often lead to an over-fragmentation of the acoustic space into clusters that are difficult to interpret or align with domain knowledge [Stowell, 2022].

Self-supervised learning, on the other hand, leverages auxiliary or proxy tasks to learn general-purpose representations from unlabelled audio. While this approach has revolutionised speech and music processing, its performance in bioacoustics remains limited [Kather et al., 2025].

Few-shot learning, as described previously, also remains a significant challenge. The high intra- and inter-species variability in vocalisation patterns, coupled with recording heterogeneity, makes it difficult for few-shot models to generalise reliably. As a result, current approaches remain far from providing a definitive solution for the majority of bioacoustic tasks.

Particularly for AIID, these limitations reveal that the machine learning methods cannot yet deal with the level of complexity that multispecies AIID presents. Addressing AIID at scale requires methods that can learn from limited and heterogeneous data and generalise across species. This development requires the curation of comprehensive databases for AIID.

Chapter 3

Few-shot bioacoustics: Rethinking the big-data paradigm.

Before delving into the main topic of this thesis, it is useful to reflect on how computational bioacoustics, particularly when approached with machine learning (ML), is typically framed.

ML implies the ability to automatically learn patterns from data. In many domains like computer vision and speech recognition, early success with ML and deep learning (DL), are due to the availability of large-scale labelled datasets and well defined and broad tasks. However, as research questions become more refined, the assumption that large, homogeneous datasets can be assembled to support each new task begins to break down. At that stage, the limitations of the traditional supervised learning paradigm become evident.

Computational bioacoustics has followed a similar trajectory. Initial work often focused on broad coarse tasks such as detecting animal sounds or identifying species presence. These tasks could be effectively addressed using supervised learning pipelines trained on moderately sized datasets. However, current bioacoustic research is moving towards more complex and nuanced goals, such as distinguishing call types, recognizing individual animals, determining sex and developmental stage from vocalizations, and operating under real-world, unconstrained conditions. These tasks present important challenges to the traditional DL approach heavily dependent on the availability of large annotated datasets. First, the data required to train models for such detailed questions is often difficult or expensive to annotate, and second, many phenomena might be so rare that makes it impossible to register in sufficient quantity. The resulting

models, when trained on narrow, highly specific datasets as such, will tend to have extremely limited applicability.

Importantly, many bioacoustic tasks like individual identification, species classification or call detection, are not singular stand-alone problems. They are better described as composed of many fine-grained, small-scale subtasks, that present a great diversity of acoustic variability across species, environments, and recording conditions. This makes the application of supervised learning particularly challenging, not just because annotated datasets are limited, but because the range of conditions a model needs to generalize across is too vast to be exhaustively represented during training. In other words, the supervised learning paradigm has reached a fundamental limitation: its reliance on large, task-specific datasets hinders the development of truly generalizable models.

Often, a common response to this problem is to build highly specialized models tailored to a specific task, species, or dataset. While these models may perform well under the narrow conditions they were trained on, they typically fail when applied more broadly. This makes them difficult to scale or deploy in real-world bioacoustic settings, and limits their usefulness in addressing broader biological questions.

This chapter provides a framework for rethinking bioacoustic tasks not as single, large-scale learning problems, but as a fragmented collection of small-scale tasks. We argue that addressing bioacoustic challenges requires moving beyond the big-data paradigm and embrace learning strategies such as few-shot learning, that are better suited to operate in an inherently fragmented scenario. This perspective better supports the development of general-purpose systems that are capable of adapting to different species, tasks, and recording conditions. Additionally, this framing also helps justify the design choices made later in this thesis, particularly in relation to multispecies AIID.

While today this framing may seem more evident, it is important to recall that a few years back, computational bioacoustics was only beginning to move beyond traditional supervised learning pipelines. In 2021, we launched the few-shot bioacoustic event detection challenge [**?**], which has shown to be a relevant effort to push for the required shift in perspective. The public challenge was intentionally designed to reflect the fragmented and heterogeneous nature of real-world bioacoustic tasks, where participants were asked to build systems able to detect events across multiple taxa, environments, and recording setups. The goal was not only to create a challenging illustration of real-world bioacoustic tasks, but also to encourage the development of unified approaches capable of handling high diversity without relying on highly specialised and narrow models.

This chapter outlines the framework designed for the public challenge and draws parallels to the AIID problem. In Section 3.1, we further argue for a shift

in perspective: from treating bioacoustics as a big-data domain to understanding it as a diverse and fragmented collection of small-scale learning problems. The section ends with an introduction of alternative methods to SL that better support this scenario. Section 3.2 describes the few-shot framework and the structure of the public challenge, with particular emphasis on the heterogeneity of the datasets and the generalisation demands for participating systems. Through the analysis of system performance across conditions, we are able to identify the key challenges faced when dealing with such heterogenous landscapes and that limit the systems' generalisation abilities. These important factors are discussed in Section 3.3, and we conclude in Section 3.4, where we look at the potential implications these exploration has on AIID. By drawing comparisons between the multispecies AIID task and the few-shot bioacoustic event detection task, we further motivate and justify the approach to AIID.

The work presented here is adapted from the following publication:

- Nolasco, Ines, et al. "**Learning to detect an animal sound from five examples.**" Ecological informatics 77 (2023): 102258. Nolasco et al. [2023b]

3.1 Approaching bioacoustics as a collection of small-scale tasks

There are two main aspects that drive the definition of bioacoustics as inherently fragmented. First, bioacoustics aims to study the sounds of animals across very different taxonomic groups. The vocal production systems of these animals vary widely, and thus the vocalisations they produce exhibit very different acoustic characteristics.

Acoustic variability also arises from environmental differences, habitat types, recording conditions, and hardware differences. Also different recording strategies such as handheld microphones, passive monitoring arrays, on-body sensors, and more, introduce their own set of very specific acoustic artefacts. Consequently, we see bioacoustics as resembling more tasks in the machine listening domain (e.g., everyday sound classification), rather than other seemingly related fields such as human speech recognition.

The second driver of this framing is the aim to build general-purpose systems that can effectively adapt to new domains and operate beyond the specific environments and species they were trained on.

This generalization-oriented perspective has implications beyond performance and scalability. Bioacoustic systems that adapt across species and tasks may

also help uncover shared acoustic structures or biological principles, offering insights into evolutionary processes. Thus, systems that generalize well support both practical applications and scientific research.

Despite the need for a unified framework, bioacoustics is often separated into narrow domains, with rigid, highly specialized models. For instance, systems developed for underwater mammals are rarely transferable to terrestrial ones. Similarly, bat detection systems trained on ultrasonic sounds do not generalize to other taxa [Van Merriënboer et al., 2024]. Changing task granularity also creates challenges: systems trained on coarse labels like species may fail to make finer distinctions, such as individual identity.

The key question addressed in this chapter is how to operate in such a fragmented landscape while maintaining the ability to generalize and adapt to new domains. The pursuit of ever-larger datasets and rigid supervised paradigms has reached a clear limit point.

Instead, we propose approaching computational bioacoustics through a few-shot learning paradigm, (see details in 2.3.6). Approaches to few-shot learning can be split into two main categories: transfer learning and meta-learning [Schaul and Schmidhuber, 2010]. They differ mainly in how they leverage prior knowledge to perform domain adaptation.

Transfer learning methods first train a model on a large dataset of base classes and then fine-tune it on a few examples from novel classes [Parnami and Lee, 2022]. However, fine-tuning with very few examples can easily lead to overfitting. To address this, methods like SimpleShot [Wang et al., 2019] use a pretrained feature extractor and rely on embedding-based nearest-neighbour classification in a normalized space. Similarly, Chen et al. [2020c] apply cosine-similarity based classification using embeddings from pretrained backbones.

This approach aligns with the idea of using pretrained models as encoders to extract general-purpose features. By learning robust representations from a diverse set of base classes, these encoders can produce embeddings that are relatively task-agnostic, i.e., they are not tightly tied to a specific output layer or label space. While the embeddings still reflect the biases of the pretraining data and task, they often capture general acoustic patterns that can transfer to a range of downstream tasks. This allows models to perform classification on the new domain, without requiring full retraining of the feature extractor.

In contrast, meta-learning trains on a distribution of tasks, rather than a single task. Its goal is to learn learning strategies that generalize across tasks. Instead of optimizing learning for a single task, the model learns how to learn (Meta learner). This training process, where the model is exposed to multiple tasks, promotes quick adaptation to new tasks at test time with very little data. This idea of training a model over multiple tasks, is often implemented

through episodic training. Instead of training on a global label space, the model is exposed to a sequence of episodes, each simulating a small classification task composed of a support set (few labelled examples) and a query set (unlabelled examples to classify). This setup aligns the training objective with the few-shot scenario encountered at test time. Episodic training is particularly associated with metric-based methods like **prototypical networks** (see Sec.2.3.6), where the model learns an embedding space and classification is performed by computing distances between query examples and class prototypes (i.e., the mean embedding of the few support examples per class).

In this way, few-shot learning, whether through transfer learning or meta-learning, offers a natural framework for tackling the heterogeneity of bioacoustics. It enables flexible models that can operate in the small-data regime while still generalizing across tasks, species, and recording contexts. Crucially, it allows us to go beyond rigid supervised learning assumptions, enabling more scalable and biologically meaningful approaches to animal sound analysis.

3.2 Few-shot bioacoustic event detection: The public challenge

The Few-shot Bioacoustic Event Detection task was organised as part of the broader DCASE challenge, which each year hosts a variety of tasks in the fields of Acoustic Scene Analysis and Sound Event Detection [1]. The task was designed to illustrate realistic bioacoustic tasks through a few-shot learning framework and it ran for four consecutive years (from 2021 to 2024) in the presented format. In this section, we describe the set up, datasets, proposed solutions, and the evaluation metrics adopted. Additionally, we report and discuss the results of participating systems from the 2022 and 2023 editions of the challenge. Given the evaluation datasets for both years were highly similar, this allowed us to conveniently analyse and compare results across overlapping data, providing a more consistent assessment.

We formulate the few-shot bioacoustic sound event detection task is as follows:

Given one long audio recording (or multiple audio recordings), as well as annotations on the onset and offset time for each of the first five sound events of interest, identify the onset and offset times for all other such sound events in the recording(s) (Figure 3.1a).

To support training of these systems we make use of multiple bioacoustic datasets (Figure 3.1b) representing a range of taxa and recording conditions,

[1] `https://dcase.community/challenge2021/`

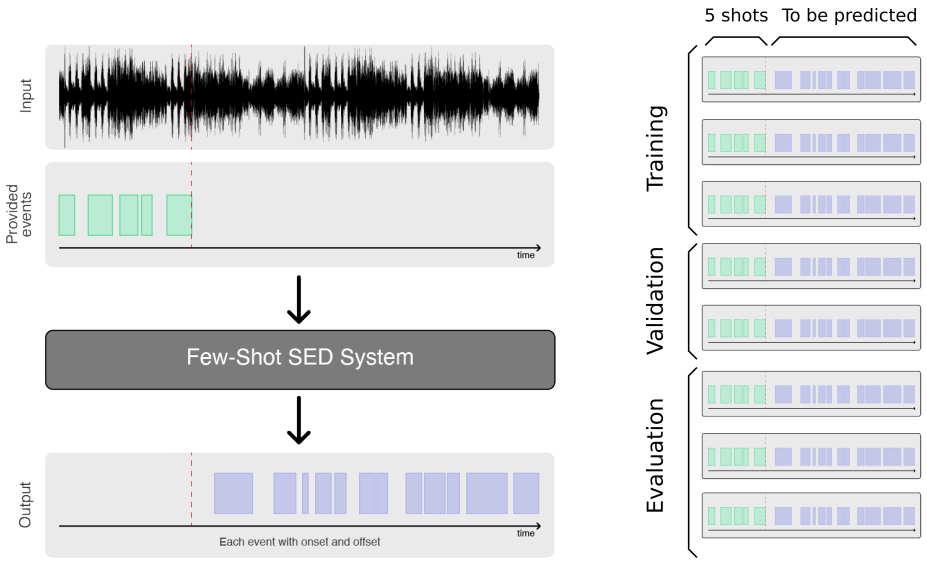


Figure 3.1: (left-side) Few-shot sound event detection: the first 5 sound events are given as examples—in standard supervised learning they would be considered the training set—and the remainder must then be detected. (Right-side) Few-shot sound event detection as a *meta-learning* problem. Each of our datasets represents a different but related few-shot task. The overall goal is to use the training and validation datasets collectively to train or otherwise develop a system that, when presented with 5 sound events from any of the evaluation datasets, can perform well at detecting the remaining events.

each annotated with a different target sound category. Note that we do not consider multiple classes in one dataset, each dataset represents a single-class problem. Other sounds are undoubtedly present in almost all audio recordings, but these are considered to be background noise (clutter/distractor events).

We choose few-shot rather than one-shot learning because animal sounds of interest often cover a range of variability: for example, there may be multiple call types in the set of sounds of interest, or calls from multiple distinct individuals within a group. Five as the number of examples is a conventional choice in few-shot.

Note also that we choose to use the *first* occurring events as examples, rather than a randomly-selected set. This reflects typical practice in bioacoustics, in that acoustic data are typically labelled in contiguous time segments which may or may not be fully representative of the entire data set, and should be tractable for future users of few-shot acoustic systems. It offers one benefit, that an algorithm may make use of the strong assumption that the periods between the first five examples fall into the negative class. It also aligns with common scenarios, such as manually labelling data during a pilot phase and then deploying a recognition system to automatically label incoming data. However,

it also presents a risk: the first sequence of examples may be similar to each other in some way which is not representative of the whole set of events, for example if the acoustic environment or animal behaviours exhibit non stationary characteristics but change over time.

3.2.1 Datasets

A conventional machine learning experiment uses a single dataset partitioned into three subsets, used for training, validation, and evaluation (test). In the few-shot formulation, we also divide the data into these three partitions (training/validation/evaluation), but each in fact contains *multiple datasets*, and each dataset represents one example of a few-shot task. Within each dataset, there are one or more audio files, each accompanied by a CSV text file giving the start time, end time and label of the targeted audio events. The label can be POS for a positive example of the target class, NEG for a negative example (background or non-target sound event), or UNK for unknown cases, where the human annotator(s) were unsure whether a sound event should be considered in the positive class. Such UNK cases may often occur in complex wildlife sound scenes; our chosen strategy was to explicitly label these cases, allowing algorithm designers to make their own decisions on how to handle them, but to exclude UNK time-regions from the evaluation measures (described later) since their correct label is ambiguous. For each dataset, the first five POS events are the "few shots" from which the rest should be inferred.

A *development set* is provided when the challenge is launched, consisting of the predefined training and validation sets to be used for system development. The development set consists of datasets from multiple sources with audio recordings and associated reference annotations in our specified format. More specifically, for the training set multi-class temporal annotations were provided for each recording (with multiple POS/NEG/UNK columns in the data, one per class), while for the validation set single-class temporal annotations (POS/UNK) were provided for each recording.

A separate *evaluation set* was kept for evaluating the performance of the systems. During the task, only the five POS event annotations were provided for each of the recordings for the class of interest. The developed systems had to use those five annotated events and then learn to detect the same type of events throughout the rest of the recording. The true annotations for the rest of the recording were kept private for evaluation.

Table 3.1 presents an overview of all the datasets in the development and evaluation sets used in the 2022 edition of the challenge. To summarise, together these datasets represent some of the wide variety of bioacoustic sound event

detection tasks, and were selected to give broad coverage of some of the key axes of variation, such as rate of occurrence of the target sound, length of calls, background noise (SNR), taxa, etc. Some of these quantitative characteristics are summarized in Table 3.1. A visual representation of each dataset is presented in appendix in Figure **??**. Further description and analysis of these datasets are presented in appendix **??**, including spectral and temporal profiles for both target sounds and background soundscapes.

2022	Dataset	Taxon	Mic type	# Files	Total duration	# Labels	# Events	Density	Mean event length (sec)
Training	BV	Birds	fixed	5	10 hours	11	9026	0.038	0.15
	HT	Mammals	on-body	5	5 hours	5	611	0.047	1.42
	MT	Mammals	on-body	2	70 mins	4	1294	0.042	0.15
	JD	Birds	on-body	1	10 mins	1	357	0.062	0.11
	WMW	Birds	various	161	5 hours	26	2941	0.25	1.54
Val.	HB	Insects	handheld	10	2.4 hours	1	712	0.67	11.67
	PB	Birds	fixed	6	3 hours	2	292	0.003	0.11
	ME	Mammals	handheld	2	20 mins	2	73	0.011	0.19
Evaluation	CHE	Birds	fixed	18	3 hours	3	2550	0.263	1.94
	DC	Birds	fixed	10	95 mins	3	967	0.350	1.66
	CT	Mammals	on-body	3	48 mins	3	365	0.017	0.16
	MS	Birds	fixed	4	40 mins	1	1087	0.084	0.18
	QU	Mammals (marine)	on-body	8	74 mins	1	3441	0.045	0.06
	MGE	Birds	fixed	3	32 mins	2	1195	0.194	0.27

Table 3.1: Information on each dataset used in the 2022 challenge. "Density" is calculated as in signal processing: (total duration of events) / (total duration of audio), thus values close to 0 are sparse, and close to 1 are dense.

The datasets represent diverse challenges for the few-shot event detection systems that are trained and evaluated on them. For each dataset, the provided 5 events are used to specify the class of target sounds. The extent to which a small set of calls can be representative depends on various factors such as vocabulary size and how stereotyped the calls are. We measure stereotypy in the data by computing similarity between events, both between the selected five initial events and also between these and the remaining events. The analysis indicates that this characteristic varies considerably for each dataset, demonstrating that for certain tasks events are more stable across time than others. (see appendix **??**). Together with sound to noise ratio and sparsity/density of call events, stereotypy is expected to be one of the key contributors for the variation in performance across datasets.

During the curation of this dataset, several deliberate choices were made to ensure that the data could effectively support the development of few-shot learning systems. These design decisions are important to describe, as they can

introduce biases and influence downstream results and analysis.

As described above, across the various tasks captured in this dataset, some recordings are densely populated with target events, while others are relatively sparse. We intentionally did not select recordings based on this characteristic, in order to better represent the natural recording conditions in which the sounds were captured. However, we required a minimum of ten positive (POS) events per recording, following the initial five provided examples. This ensures that participants have at least ten target-class events to detect among other non-target ("distractor") sounds present across the recording. The inclusion of distractors is essential to assess whether systems are truly learning discriminative characteristics of the target class. Regarding recording length, we favoured audio files up to approximately 1 hour to facilitate data loading and processing, although no strict constraint was imposed.

Annotations were performed by domain specialists for each dataset seprarately, (see Appendix **??**). This implies that annotation strategies are not fully standardised across all tasks. Even if such consistency were possible, defining the onset and offset of animal vocalisations is inherently imprecise and can depend on the annotator's methodology, such as using a spectrogram for visual guidance or marking events purely by auditory perception. It also depends on the nature of the vocalisations themselves, as some calls have gradual or ambiguous beginnings that are difficult to define precisely. In this work, the impact of these heterogeneous annotation strategies is mitigated in two ways: (1) by using an evaluation metric that considers a prediction correct if it overlaps with a sufficient proportion of the ground-truth event, and (2) by treating each recording and its five provided examples as an independent task, even when belonging to the same dataset. This approach encourages systems to learn to mimic the annotation strategy specific to each recording. Thus, it is important to note that models do not necessarily learn an explicit concept of where an animal vocalisation begins or ends, but rather reproduce annotation patterns.

Finally, when target events overlap with other vocalisations, the annotation protocol treats these mixtures as a single UNK (unknown) event. As described in the evaluation section, UNK events are not required to be detected so these do not count as missed detections, nor false detections. This ensures that system performance is not penalised by the presence of overlapping vocalisations.

3.2.2 Baseline systems

Upon the launch of the challenge, two base solutions were proposed. These represent standard methods that can be used to compare participating systems against. One is an approach commonly used in bioacoustics based on spec-

trogram cross-correlation and the other is a deep learning approach based on prototypical networks.

Template matching (cross-correlation) is a signal-processing method that works by cross-correlating waveforms or spectrograms. In this implementation we use the scikit-image's `match_template` function applied to power spectrograms. This function computes normalised cross-correlation to find instances of a template in an image, returning values ranged between -1.0 and 1.0, with higher values corresponding to higher correlation. Our few-shot template matching method computes cross-correlation across the time axis between each of the events (shots) provided for a file and the rest of the recording. A different detection threshold is set for each audio file based on the max value of the cross-correlation results between the shots provided. Peak picking is performed on the results of the template matching algorithm, with any peak above the threshold corresponding to the centre of a detected event in that recording. Borders of the predicted event are assumed to align with the beginning and end of the template when it matches. Each of the 5 templates is used separately for matching, and the resulting event predictions are collapsed into a single binary prediction vector which will produce the final events predicted for the class of interest.

Template matching can work well in noisy audio, providing the target signal is acoustically (a) distinct from the background sounds and (b) stereotyped, i.e., not strongly varying in character. We thus expect template matching to work well in some of the scenarios we study, but to perform very poorly in others.

Prototypical networks is a deep learning technique that classifies inputs based on their distance to class prototypes, averages of embedded support examples, (see further details in Section 2.3.6). The model is trained using *episodic training*, where each episode mimics a small classification task with a limited number of examples per class.

In our setup, we adopt an episodic training setup as 2-way 5-shot configuration when evaluating a single sound event type, treating the two classes as active versus inactive. During evaluation, we follow a binary classification strategy inspired by Wang et al. [2020b]. The first 5 positive annotations in a file are used to compute a prototype for the target class, while the rest of the file is assumed to contain mostly negative regions. Negative prototypes are generated by randomly sampling regions assumed to be inactive. Each query sample is predicted to have the target sound

active if its embedding is closer to the positive prototype than the negative. This process is repeated 5 times with different negative samples and the final prediction is the average of these iterations. To reduce false positives, we discard any predicted event shorter than 60% of the shortest shot in the file.

3.2.3 Evaluation

For the evaluation of this task, we employ an event-based F-measure with macro-averaged metric, to evaluate the match between true and predicted events. The main complexity is related to the detection of a match between ground truth events and predicted events. Traditional approaches use onset detection based metrics and fixed-size evaluation windows [Mesaros et al., 2019]. Given the great variation between datasets and characteristics of the events we want to detect in this task, these approaches are not suitable. Instead, we use the Intersection over Union (IoU), with 30% minimum overlap to produce a list of possible matches of the predictions. The IoU metric measures the overlap between two time intervals. Given a predicted event interval P and a reference event interval R, the IoU is defined as the ratio of the length of their intersection to the length of their union:

$$\text{IoU}(P, R) = \frac{|P \cap R|}{|P \cup R|}, \tag{3.1}$$

where $|P \cap R|$ is the duration of the overlap between the predicted and reference events, and $|P \cup R|$ is the total duration covered by either event. An IoU threshold of 30% means that a predicted event is considered a match if at least 30% of the combined interval is overlapping.

The result of applying IoU is a list of predicted events that are candidate matches. For each ground truth event, a single best match is selected by applying the Hopcroft-Karp-Karzanov algorithm for bipartite graph matching, a similar procedure as used in the `sed eval` toolbox.[2]

In a sound event detection (SED) task we can define True Positives (TP) as predicted events that match ground truth events, False Positives (FP) as predicted events that do not match any ground truth events, and False Negatives (FN) as ground truth events that are not predicted. In this task, ground truth events consist of POS events of the class and UNK events that have some uncertainty associated to the assigned class. The UNK label is typically assigned if the annotator is not sure if the event belongs to the class of interest, in other situations the annotator knows it is not the class of interest however the event

[2] `http://tut-arg.github.io/sed_eval/generated/sed_eval.util.event_matching.bipartite_match.html`

lookssounds like it could be. This decision is thus very subjective and dependent on each dataset. For evaluation purposes, matches to UNK events do not count towards calculating TP, FP or FN. In doing so, we ensure that the subjectivity associated with assigning the UNK label does not impact the performance score of the systems.

The procedure we employ is:

1. Apply IoU and bipartite graph matching between predicted events and ground truth POS events only, resulting in TP.
2. Apply IoU and bipartite graph matching between remaining predicted events, that did not match with any POS event, and ground truth UNK events only.
3. Compute FP as the number of predicted events that were not matched to either POS or UNK events.
4. Compute FN as the number of POS ground truth events that were not matched by any predicted event.

This is applied to each dataset in the evaluation set where we compute the F-score metric, (in Sec.2.3). The reported results are the harmonic mean over all the datasets, which is appropriate for combining percentage results, and ensures that a system should perform well across all datasets to achieve a strong score.

We thus use an averaged F-score as our main summary statistic for each submitted system. To explore system performance in more detail, we also inspect the F-scores per dataset, and per class in each dataset, in particular to examine whether differences in acoustic characteristics correlate with differences in performance.

The F-score metric is designed to summarise how well a system's outputs correspond to the desired outputs. However, there are many factors that affect the usefulness of such outputs, meaning that it is difficult to estimate a technology readiness level from only numerical scores. Hence, in addition to our quantitative analysis, we conduct a qualitative user-oriented analysis of selected system outputs, gathering feedback from expert users (annotators of the datasets).

3.2.4 Results

In the 2022 edition, 15 teams participated submitting a total of 46 systems and in 2023 there were 6 teams with a total of 22 systems. We present in Table 3.2 the overall scores of the best system submitted by each team in these two editions of the challenge. The challenge can be seen to be a difficult one: the baseline systems, and many teams, obtained F-score averages below 25%. On

Team	**Evaluation (95% CI)**	**Validation**
Du_NERCSLIP_23 [Yan et al., 2023]	**61.83** (61.23-62.32)	75.6
Du_NERCSLIP [Tang et al., 2022]	60.22 (59.66-60.70)	74.4
Liu_Surrey [Liu et al., 2022a]	48.52 (48.18-48.85)	50.03
Martinsson_RISE [Martinsson et al., 2022]	47.97 (47.48-48.40)	60
Hertkorn_ZF [Hertkorn, 2022]	44.98 (44.44-45.42)	61.76
Liu_BIT-SRCB [Liu et al., 2022b]	44.26 (43.85-44.62)	64.77
Wu_SHNU [Wu and Long, 2022]	40.93 (40.48-41.30)	53.88
XuQianHu_NUDT_BIT [Liu et al., 2023]	37.71 (36.98-38.23)	63.94
Moummad_IMT [Moummad et al., 2023]	37.32 (36.82-37.74)	63.46
Zgorzynski_SRPOL [Zgorzynski and Matuszewski, 2022]	33.24 (32.69-33.69)	57.2
Gelderblom_SINTEF [Gelderblom et al., 2023]	26.79 (26.13-27.29)	36.6
Mariajohn_DSPC [Mariajohn, 2022]	25.66 (25.40-25.91)	43.89
Jung_KT [Lee et al., 2023]	23.74 (23.14-24.17)	81.52
Willbo_RISE [Willbo et al., 2022]	21.67 (21.32-21.97)	47.94
Zou_PKU [Yang et al., 2022]	19.20 (18.88-19.51)	51.99
Huang_SCUT [Huang et al., 2022]	18.29 (18.01-18.56)	54.63
Tan_WHU [Tan et al., 2022]	17.22 (16.82-17.55)	54.53
Li_QMUL [Li et al., 2022]	15.49 (15.16-15.77)	47.88
Wilkinghoff_FKIE [Wilkinghoff and Cornaggia-Urrigshardt, 2023]	13.31 (12.83-13.67)	62.64
baseline-TempMatch [Morfi et al., 2021]	12.35 (11.52-12.75)	3.37
baseline-ProtoNet [Morfi et al., 2021]	5.3 (5.1-5.2)	28.45
Zhang_CQU [Zhang et al., 2022b]	4.34 (3.74-4.56)	44.17
Kang_ET [Kang, 2022]	2.82 (2.76-2.87)	-

Table 3.2: F-score results (in %) per team (best scoring submission) on 2022 evaluation and validation sets. Systems are ordered by higher scoring rank on the evaluation set. These results and technical reports for the submitted systems can be found on task 5 results page [DCASE, 2022] and [DCASE, 2023].

the other hand, methods could be designed which reach well over 40% F-score average, and up to 60% (Table 3.2). Such performances were much stronger than expected based on the task difficulty and the results on the first edition of the task (2021).

Several teams adopted prototypical networks, likely influenced by the baseline. Improvements over the baseline were achieved via data augmentation, intelligent post-processing, and better construction of the negative prototypes (Liu_Surrey, XuQianHu_NUDT_BIT, Jung_KT, Wu_SHNU, Jung_KT, Willbo_RISE). Some systems also employed *transductive inference*[Boudiaf et al., 2021], which adapts the feature space at test time using information from the test input, (Liu_Surrey, XuQianHu_NUDT_BIT, Li_QMUL, Tan_WHU, Zou_PKU).

The top-scoring systems,(Du_NERCSLIP_23 and Du_NERCSLIP), used frame-level embeddings achieving high time resolution. This is particularly beneficial for short-duration classes like QU (see Figure **??** for a breakdown of F-scores by dataset). In 2023, this approach was combined with multi-task learning, including voice activity detection, yielding a 2% score increase.

The next system in the ranking is Liu_Surrey, which implements a novel approach designed to optimise the contrast between positive events and negative prototypes. This, together with an adaptive segment length dependent on each target class, works well across all the evaluation sets.

The problem of adapting to different lengths of events across target classes was also directly addressed by other submissions. Both Martinsson_RISE and Zgorzynski_SRPOL implemented an ensemble approach where each individual model focuses on a different input size range. In Liu_BIT-SRCB this is explored through a multi-scale ResNet, and in Willbo_RISE with a wide ResNet containing many channels. Also in XuQianHu_NUDT_BIT, they implement a novel adaptive mechanism - squeeze/excitation block - designed to assign different weights to different channels of the feature map.

Other strong submissions looked at the process of building the negative prototype, for instance Liu_Surrey optimized contrast between positive and negative embeddings.

Importantly, the submitted systems exhibit large variations in their performance across datasets, and also intra datasets at the target class level. Figures **??** and **??** presents the F-score results for the various datasets and per target class in the multi-class datasets. The easiest classes to be detected are CHE_chaffinches, CT_chirpgrunts and DC_robins, where several systems reach above 75% F-score. On the other side, CT_Chitters, DC_Cuckoo and the QU_Quacks seem to be the classes where systems struggled the most to make correct predictions. This shows the challenge in developing unified systems that can address a high variety of acoustic characteristics and tasks.

Inspecting the characteristics of the methods performing most strongly in the challenge, broadly across all editions, we observe some general tendencies (we present a summary table of systems characteristics in appendix **??**). Across editions, systems mostly relied on standard Convolutional neural networks and Melspectrograms with per channel energy normalisation (PCEN Lostanlen et al. [2018]). While architectures were similar, strategies for training and inference varied: some used meta-learning with prototypical networks, others fine-tuned models with cross-entropy. There is a roughly equal balance of the two main paradigms: meta-learning with prototypical networks, versus fine-tuning or otherwise adapting a network trained using cross-entropy.

Methods to select time regions for training and pseudo-labelling were common. Pseudo-labelling, used by several top teams, involves self-labelling data to increase training volume. Successful systems also commonly used explicit methods to control the duration of the detected events. In many cases this consists of post-processing predictions to delete/merge very short events, or estimating the typical duration from the examples. Du_NERCSLIP(23) and Wolters et al. [2021] made use of neural network architectures specifically trained to infer and output region annotations.

Overall, the different approaches submitted illustrate the introduction of ideas to address challenges related to this task: how to deal with very different event lengths; how to construct a negative class when no explicit labels are given for this; and how to bridge the gap between classification and detection for few-shot sound event detection. These challenges derive from the combination of few-shot learning with sound event detection, and hence are not addressed in standard few-shot learning [Wang et al., 2020a].

3.3 Discussion

This challenge presented a few-shot learning formulation for bioacoustic sound event detection, demonstrating that both meta-learning and transfer learning approaches can be used to bioacoustic detection tasks without the need for large annotated datasets. Furthermore we showed the ability to adapt to novel tasks without extensive training. By framing the problem in this way, we address a wide variety of task conditions using a unified machine learning approach. Overall systems achieved over 30% average F-score across all tasks (Figure **??**), significantly outperforming traditional template-matching and standard deep learning methods.

Beyond variable event lengths, other characteristics of the data create important challenges for the participating systems, such as background noise, the presence of distractor events, and intra-class variability, all of which are com-

mon in real-world bioacoustic recordings. Participating systems have addressed directly some of these by developing strategies for adapting to different event durations, better prototype formation, and negative example construction.

Our evaluation shows that improved prototypical network methods create powerful embeddings, useful even with no test-time adaptation. It is impressive that a single vector space could be used to represent our diverse bioacoustic tasks. The present work on few-shot learning thus offers a different perspective on representation learning for sound in general, and animal sound in particular.

Another key finding is the effectiveness of structured embedding spaces. Many systems relied on learned embeddings and distance-based reasoning to perform detection. Prototypical networks and other metric learning approaches performed well, and even systems without explicit few-shot architectures benefited from embedding-based representations. This suggests that representation learning and distance metrics are critical components for generalization in low-data detection settings.

In summary, this chapter illustrates the practical application of few-shot learning to a real-world bioacoustic detection task. It highlights the importance of embedding learning, adaptation strategies, and careful system design to address the challenges specific to sound event detection, such as non-stationarity, duration variability, and the absence of explicit background class annotations.

3.4 Conclusion

In this chapter we motivate a different perspective of computational bioacoustics. We justify that most bioacoustic tasks cannot be described as a big-data problem. Instead, due to the high variety of acoustic characteristics, various resolution levels of the questions we pose and the general associated difficulty in acquiring and annotating data, bioacoustic tasks are better described as a collection of small-scale tasks. Operating in this landscape and aiming to build generalisable systems is challenging and requires exploring approaches beyond supervised learning.

The Few-shot bioacoustic event detection challenge was first launched at a time where this shift in perspective was starting to happen. By providing a few-shot setup with accompanying datasets and evaluation framework, we were able to not only make evident the need for alternative approaches to bioacoustic tasks, but also to create a playground where novel and unifying approaches to bioacoustics can be developed.

3.4.1 Implications for multispecies AIID task

Multispecies AIID is in its nature a prime example of the challenges described in this chapter. In short, conceptualising this problem as collection of species-specific tasks releases us from the quest for more and more annotated data. Indeed, the next chapter (4), describes the dataset compiled for our development work in multispecies AIID, where it is evident how the task fits this fragmented description of bioacoustic tasks, and also aligns with its generalisation goals.

While many of the difficulties in the public challenge are associated with it being a detection task, namely the requirement for high temporal resolution and dealing with varying event lengths, the ability to develop ML methods that can generalise to various heterogeneous tasks with very few data has direct implications on how we address the multispecies AIID classification task.

Two direct implications to AIID arise from this exploration: First, there is strong evidence that having a solid base representation of the data, even if not fine-tuned to the specific species or tasks at hand is fundamental to work within low-data scenarios. The idea of using pretrained embeddings from models more or less related to the target domain, and that were able to train on more extensive datasets is solidified in this analysis. Second, the success of prototypical networks in this context highlights the potential of distance based learning methods to learn powerful embedding spaces that can represent in a single space a multitude of varied categories. This offers a clear way forward to build unified approaches to AIID such as the ones developed in chapter 6, that promote generalisation to novel categories.

Chapter 4

A multi-species dataset for acoustic identification of individual animals

This chapter presents a novel dataset designed for developing computational methods for Acoustic Identification of Individuals (AIID) across multiple species, addressing gaps in species diversity and generalization in existing datasets.

It begins by motivating the need for such a dataset and highlighting the dataset's significance in advancing AIID research. Here, we explore key considerations such as the multi-species approach, phylogenetic relationships, and general data requirements. The chapter then details the dataset's compilation process, contents, and various challenges often associated with data collection and labelling in this context. This dataset captures a diverse range of species, from mammals to birds, emphasizing the variety in their vocalizations and conspecific recognition strategies.

Finally, the chapter presents an exploratory analysis of data representation, where we evaluate pretrained audio embeddings using Beecher Information and clustering metrics such as the Silhouette Score, to assess their suitability for multi-species AIID. This study shows that OpenL3 embeddings trained on environmental sounds offers a suitable representation of the data across hierarchical levels of the AIID task, and is expected to be a good basis for further development.

4.1 Motivation and requirements

The AIID task shares several commonalities with the few-shot bioacoustic event detection task described in Chapter 3. In both we aim to unify the approach to detect or classify examples from diverse sources and characteristics, more closely approximating the reality of bioacoustics rather than solving the task for narrow, specific conditions and species.

Developing a unified approach for AIID motivates our exploration in a multi-species scenario which contributes to the heterogeneity of sound characteristics in our study. Multi-species AIID is innovative but challenging. As existing systems focus on single-species identification, (see related work in Sec.2.2), our goal is to achieve this across all species in our dataset, without having to retrain or fine-tune models for each. In doing so, we can show the potential to generalise across species and taxonomic groups. Beyond the technical innovation, there is a concrete application potential in developing general multi-species systems for AIID, particularly in research areas like animal behaviour studies, biodiversity and wild-life monitoring. Furthermore, there is a lack of comprehensive datasets to fully develop multi-species AIID. While some small datasets have been released which contain individual ID labels[1], they often focus on single species, and predominantly birds. To create the present multi-species dataset for AIID, we have incorporated some openly available datasets, and used other dataset donations by research groups studying animal vocalisations and acoustic signatures (AS). The present dataset is far from representing the vast diversity of vocal animals, however it offers sufficient variation—ranging from birds to mammals, both large and small—to test whether the algorithms can generalise to other individuals and species.

In ML terms, AIID is framed as a classification task, not detection. Our focus is on assigning ID labels (animal ID) to isolated vocalisations rather than locating them within a recording. Each vocalisation can be associated with 3 labels: taxonomic class (mammal, bird ...), species, and animal ID.

Due to the difficulty in gathering data with labels for animal ID, the selection of the species used does not follow strict constraints. Instead we aim to use the data that is available while keeping a somewhat balanced representation between mammal and bird species. With respect to the number of examples per individual, no class balancing is enforced; however, a minimum of 10 recordings per individual is required. This decision reflects real-world bioacoustic scenarios, challenging our methods to work under adverse conditions while maintaining the supervised learning paradigm.

[1] see a comprehensive list of available datasets here: https://bioacoustic-ai.github.io/bioacoustics-datasets/

The process to obtain animal ID labels varies depending on the species and its living conditions. If the animals live mostly alone and stay within the same location, the animal ID can be discerned from the location of the recording alone. A similar situation occurs for recordings made from laboratories in which it is possible to isolate the animal or to have a precise location of the sound source. However, for many other animal species, obtaining such labels and recordings of individuals requires an expert person who can track and follow the animals while live annotating the recordings. This person can use visual cues or direction of sound to distinguish between individuals and register their id. Semi-automatic methods exist for very specific scenarios. For instance, in [Lehmann et al., 2022] Spotted Hyena vocalisations are obtained from collar-mounted recorders and distinguished between focal (from the collared animal) and non-focal calls based on loudness, cross-referencing with collar ID to identify the vocalising individual. Animals with specific behaviours, such as Storks or Manx shearwaters that mate for life and return to the same nest annually, can potentially be identified based on nest/burrow ID and sex determination. Datasets tracking individuals across different locations, weather conditions, and seasons are rare but invaluable [Stowell et al., 2018]. These datasets are crucial for testing algorithm generalisation and acoustic signature stability across varied conditions, though they are extremely challenging to compile.

The multi-species nature of our dataset opens up possibilities for exploring taxonomic relationships within the data. Specifically, we can investigate how mechanisms for vocal production, especially those encoding individual identity in vocalisations, might follow a phylogenetic process. Evolutionary biology has long recognised that closely related species often share physical similarities. This principle extends beyond morphological traits to behavioural patterns, including acoustic communication. Vocalisations are influenced by both vocal anatomy and behaviour, which can have strong genetic components, (we discuss expectancy of hierarchical traits in animal vocalisations in Section 2.1).

Hierarchical structures [Silla and Freitas, 2011], also emerge at the label level. The 3 labels (taxon, species, ID) associated with each vocalisation are not independent of each other, each label represents a different level of broadness, and knowing a label at the narrower level (ID) automatically determines the levels above (species and taxon).

Including phylogenetic aspects into AIID means that genetically similar animals might produce comparable AS, allowing us to enhance individual prediction based on its phylogenetic relationships. By analysing how individual identification features in vocalisations vary across the phylogenetic tree, we might gain insights into the evolution of individuality in animal communication. Incorporating phylogenetic information into AIID algorithms could improve their

performance, especially when dealing with species not well-represented in the training data.

Summarising, the guiding requirements for the compilation of this dataset are the following:

1. Examples are recordings of single vocalisations (no multiple or overlapping calls).
2. For each species several examples from multiple individuals are required.
3. Species should cover both mammal and bird taxonomic groups.
4. Each recording is associated with a label that uniquely identifies the vocalising animal.

4.2 Dataset description

The dataset consists of short recordings of individual vocalizations from 105 individuals representing 8 species, including 3 mammal species and 5 bird species. Table 4.1 and Figure 4.1 provide an overview of the dataset, highlighting the distribution of recordings across species and taxonomic groups.

The unique value of this dataset lies in the diversity of calls and acoustic characteristics across the different species. This diversity is visually demonstrated in Figures 4.2 and 4.3, which showcase spectrograms of vocalizations from three individuals for each species included in the dataset.

Species	Taxon	Number of individuals	Total num. of items	Av. num. items per ID
Chiffchaffs	Birds	23	6762	294
Tree pipits	Birds	10	1337	133.7
Little owls	Birds	16	952	59.50
Peafowls	Birds	18	360	20
Eurasia eagle owls	Birds	7	463	66.14
Hyrax	Mammals	19	1162	61.16
African lions	Mammals	5	164	32.80
Grey wolves	Mammals	7	268	38.29

Table 4.1: Summary of the multispecies dataset. For each species, the corresponding taxonomic group, the number of individuals, total number of recordings, and the average number of recordings per individual are reported.

4.2.1 Structure

To make the data directly usable for machine learning, the dataset is organised into development (train and validation sets) and evaluation sets. Table 4.2

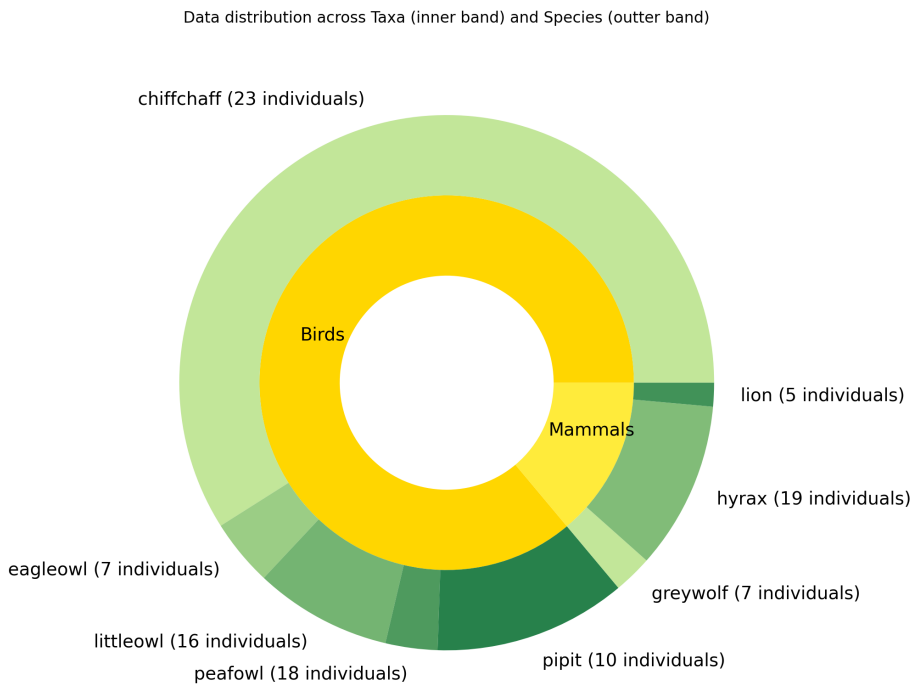


Figure 4.1: Visualisation of the species distribution in the dataset. Inner band represents the amount of recordings in the two taxonomic groups Mammals and Birds. The outer band represents the distribution of audio recordings across the various species. The number of individuals from each species is indicated in between brackets.

summarises the contents of the different sets. We build a suite of 3 evaluation sets to represent the scenarios in which we desire to assess the performance of our methods:

Test Set Contains examples of the same IDs as in the development set. This is the traditional ML test set where we measure how well the methods can classify new examples from the same classes on which they were specifically trained.

Unseen IDs Test Set Contrary to the **Test Set**, the examples in this set are from different IDs than those defined in the development set. The evaluation purpose is connected to the open world classification goal, where we want to measure how well the methods can make predictions on new individuals for which they have never been trained. The construction of this set follows a simple procedure: for each species used for development (predetermined as described below), we randomly select 2 individuals. Examples from these selected IDs are removed from the main pool of data

from which the development and test sets are obtained.

Unseen Species Test Set Similar to the Unseen IDs Test Set, this consists of examples of individuals from a novel species. Measuring the performance of the methods in this setting can indicate how well they perform at the taxonomic level, and also if they are capable of distinguishing at a very fine level (ID) when both ID classes and the species has never been seen. The unseen specie is Peafowl.

The dataset creation process begins by defining the Unseen IDs and Unseen Species test sets, removing their examples from the main pool of data. Next, the test set is created by randomly selecting a minimum of 10 examples for each ID from the main pool. The remaining data is then equally split into 5 folds ensuring stratification across all known IDs and species. The training set is formed by concatenating 4 of these folds, while the remaining fold serves as the validation set.

Set	**Number of items**	**Number IDs**	**Number species**
Training	6228	66	7
Validation	1575	66	7
Test	1972	66	7
Unseen ID Test	1333	21	7
Unseen Species Test	360	18	1

Table 4.2: Summary of dataset splits.

4.2.2 Species characterisation

This dataset captures a large variety of acoustic characteristics across the 8 species. Figures 4.2 and 4.3 show spectrograms of recordings for 3 individuals from each species. This visualisation is an important illustration on this variety of acoustic characteristics and levels of individual distinctiveness in the animals' calls.

All the species included in this dataset are known to have acoustic signatures and for most there are enough indication that they perform acoustic individual recognition. however the degree to which these have been studied varies. The following is a short description of the species, their vocal characteristics and the most current knowledge regarding their acoustic signatures.

Chiffchaff (*Phylloscopus collybita*)

These are small (10–12 cm) migratory songbirds inhabiting open woodlands across Europe and the Palearctic. They are insectivorous. Males are highly defensive of their breeding territory. Chiffchaff's name comes from the characteristic song, consisting of alternating "chiff" and "chaff" syllables, varying in order and length. Beyond this signature song, chiffchaffs have a complex, non-random song organisation. Song types are distinguished based on the song's syllabic content, however the duration and order in which these appear can vary [Linhart et al., 2012]. While repertoire based individual recognition has been previously established, the study in [Průchová et al., 2017] has shown that general song characteristics such as duration and interval between syllables can be used to distinguish between individuals. The recordings used here consist of bouts of territorial songs.

Tree Pipit (*Anthus trivialis*)

Migratory songbirds approximately 15 cm in length, travelling between Africa, Northern Europe, and the Palearctic. These insectivorous birds inhabit open woodlands and, like Chiffchaffs, nest on or near the ground. Males produce two main types of songs: perched songs, which are performed from elevated positions such as trees or other vegetation, and flight songs, which are performed during flight [Petrusková et al., 2008].

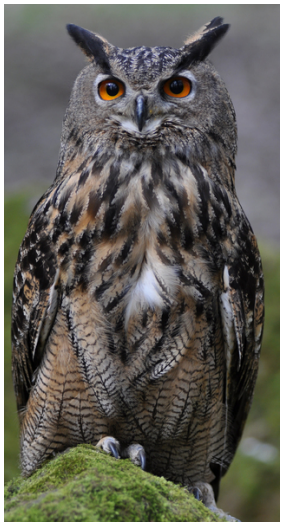
Research has demonstrated that individual Tree Pipits can be distinguished by the arrangement of syllables within their songs. The syntax of these songs is stable across seasons, enabling researchers to track individual males over time. Individual recognition of conspecifics is connected to territorial defense.[Petrusková et al., 2016]

Little Owl (*Athene noctua*)

Small (aprox. 20cm length), nocturnal raptors that are part of the Owl family. They are territorial and typically inhabit open fields. Little owls are monogamous and show high site fidelity, returning to the same territories year after year. Territorial calls are used to signal ownership, and intruders are often chased away more actively if they are strangers rather than familiar neighbours, as shown in [Hardouin et al., 2006]. Primarily "Hoot"

calls used for territory settlement and mate attraction. Other calls are used for defence against predators. Recordings include single territorial calls.

Eurasian Eagle Owl (*Bubo bubo*)



The Eagle Owl (Bubo bubo) is one of the largest owl species, found across Europe and Asia, inhabiting wetlands, woodlands, and shrublands. They are easily identified by their striking orange eyes, round faces, and prominent ear tufts. As opportunistic hunters, they prey on small mammals such as rodents and rabbits, as well as birds and occasionally larger mammals.

These owls are non-migratory and highly territorial, often maintaining lifelong pair bonds and exhibiting nest fidelity by returning to the same nest annually [Grava et al., 2008]. Vocal communication plays a central role in their lives, serving functions such as mate attraction, territorial defense, and pair bonding. They frequently use prominent topographic features such as rocky pinnacles, stark ridges, and mountain peaks as regular song posts. These landmarks, located along the outer edges of their territories, are visited regularly but only for a few minutes at a time.

Research shows that Eagle Owls possess individual vocal signatures, particularly in territorial and pair-communication calls, which remain stable over time [Lengagne et al., 2001, Grava et al., 2008]. These individual characteristics are primarily associated with temporal call patterns and the ascending part duration, which are crucial for individual recognition. Interestingly, vocal distinctiveness is influenced by environmental conditions. Delgado et al. [2013] demonstrated that in populations with high density or homogeneous habitats, vocal individuality is reduced, likely as an adaptive mechanism to mitigate aggression.

Peafowl (*Pavo cristatus*)

Peafowls are easily recognized for their iridescent plumage and crests and for the impressive courtship displays from males to attract females. They originate in the Indian subcontinent but can be found all over the world sometimes in semi-domesticated states. They are omnivorous ground-dwelling birds. peafowls

can live alone or in mixed sex flocks.

They emit a variety of calls connected to anti-predator behaviour and courtship displays. The most common anti-predator calls is the "bu-girk" call, which is emitted by both sexes. This is a relatively loud call that includes two elements: "bu" and "girk". The combination of characteristics of these two elements encodes individuality, particularly the duration and frequency of the "bu" part of the "bu-girk" . Identity in other calls has not been established. However, sufficient support exists that courtship calls are good indicators of mate quality [Yorzinski, 2014].

African Lion (*Panthera leo*)

Large felines (120–230 kg) inhabiting sub-Saharan savannas. Lions form prides with a complex social structure involving females, cubs, and a smaller coalition of males. These prides can be defined as fluid fission-fusion societies, meaning that individuals will often associate in smaller subgroups rather than the entire pride. Lions use roars to locate their pride members, territory maintenance, and group dynamics. Roars are low-frequency (F0 ranges between 40-250Hz) sounds that can carry over long distances [Grinnell and McCOMB, 2001]. Researchers have shown that roars contain information about the individual and lions rely on them for vocal recognition of other lions. Studies show that the temporal patterns of the F0 contour are a key feature that encodes identity [Wijers et al., 2021].

Rock Hyrax (*Procavia capensis*)

Hyraxes are small (30-70cm length), herbivorous mammals found in rocky areas of sub-saharian Africa and Middle East. They share resemblance with both rodents and ruminants however hyraxes are more closely related to elephants and manatees. Hyraxes are highly social animals, living in groups that can range in size from a few individuals to over 100 [Schneiderová et al., 2024].

They are known for their complex social structures and behaviours, supported by a complex vocal repertoire of calls and songs. Songs are mainly produced by males and typically consist in 3 elements - wails, chucks and snorts that are combined and ordered in order to create elaborate vocal displays. Functionally, vocalisations play a role in sexual behaviour, competition, and individual

identification [Demartsev et al., 2023]. Individuality in the vocalisations is encoded by a mix of temporal characteristics like song duration, number of bouts and rate of particular song elements, but also frequency characteristics like minimum fundamental frequency. The relative importance of these features for individual recognition may vary across different groups and populations [Koren and Geffen, 2011].

Gray Wolf (*Canis lupus*)

The wolf is the largest species of the Canidae family that still exists today. Native to Eurasia and North America, it has become extinct in much of Western Europe, the United States, and Mexico due to habitat loss and persecution. Wolves can live in diverse habitats, including forests, grasslands, and mountains [Sadhukhan et al., 2021].

Wolves have a complex social organization, forming packs typically consisting of a mated pair and their offspring. When young wolves reach sexual maturity, they disperse from their family packs, often travelling vast distances before joining or forming new packs. Packs maintain and defend territories that can span hundreds of kilometres, marked by scent and vocal signals. Howling, their primary long-distance vocalization, can carry over several kilometres and is used to coordinate pack members, defend territories, and deter rival packs. Wolves also produce growls, woofs, and barks for short-distance communication within the pack [Sadhukhan et al., 2021].

Individual recognition is essential for their social organization. Studies [Palacios et al., 2007, Root-Gutteridge et al., 2014] have shown that howls contain individual-specific characteristics, such as differences in fundamental frequency and amplitude, enabling wolves to identify each other.

4.3 Exploratory Analysis on Data Representation for AIID

In Chapter 2, we discussed the importance of acoustic features and data representation in computational bioacoustics. A crucial aspect of this field is identifying optimal ways to represent audio data for effective computational processing while preserving the critical information needed for specific tasks. As highlighted in Chapter 3, small datasets pose significant challenges, particularly when the data exhibits substantial variability. This issue is especially pronounced in AIID,

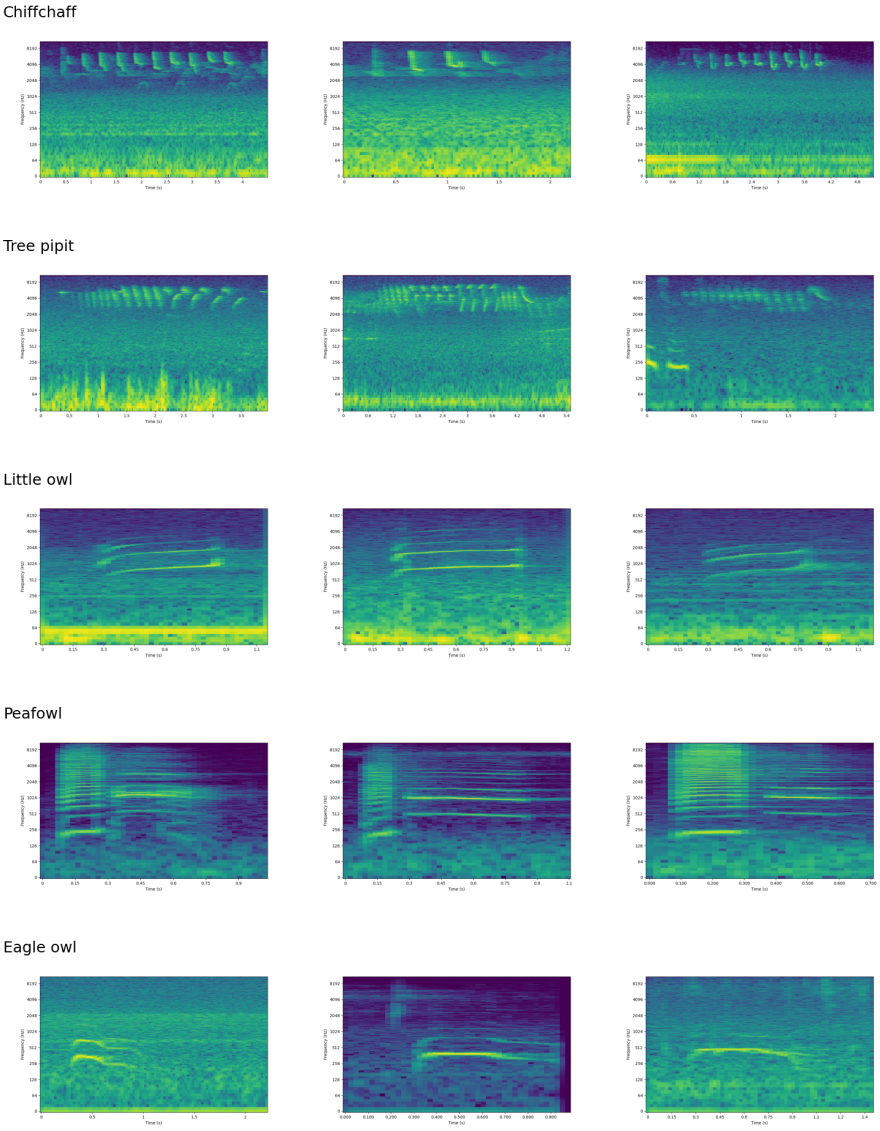


Figure 4.2: Examples of spectrograms of calls from 3 different individuals for each bird species in the dataset.

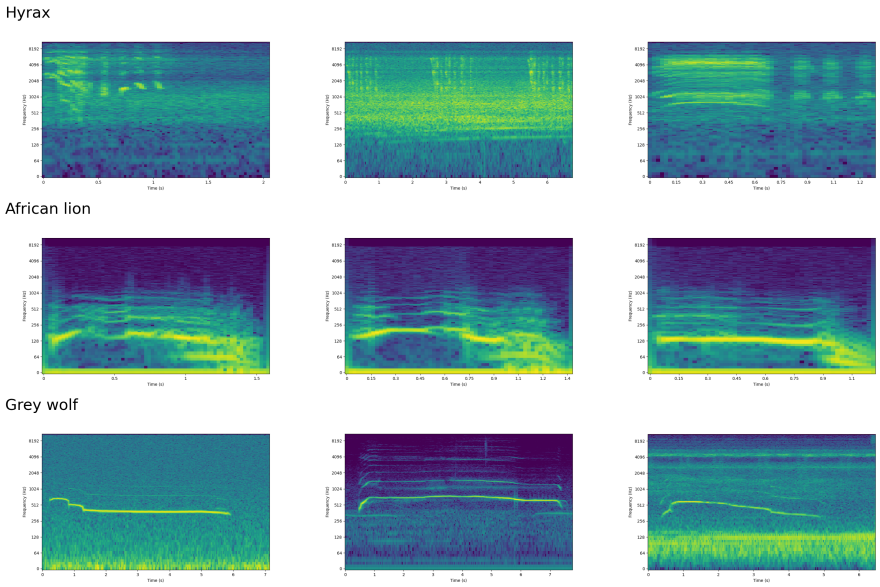


Figure 4.3: Examples of spectrograms of calls from 3 different individuals for each mammal species in the dataset.

where subtle differences in vocalizations between individuals must be discerned amidst the broader similarities shared within a species.

To process audio signals computationally, they must be digitized and transformed into compact formats that retain the information essential for the task. For AIID, the goal is to derive representations of vocalizations that encapsulate key acoustic characteristics necessary for distinguishing between individuals, as well as species and broader taxonomic groups. These representations are often informed by studies on acoustic signatures, which identify the specific features of signals that encode individuality.

However, this study focuses on developing general approaches applicable across multiple species. Consequently, we seek a unified representation capable of functioning across diverse species and acoustic signatures without requiring species-specific tailoring. This need has led us to explore the potential of large pretrained models. These models generate embeddings that represent audio signals by abstracting away irrelevant artifacts while retaining salient features. Pretrained embeddings, derived from models trained on extensive and diverse datasets, offer a promising avenue for generalization. They enable knowledge transfer, yielding representations suitable for a wide range of related tasks, including AIID. Nevertheless, a fundamental question arises: Do these embeddings preserve the information necessary for distinguishing between individuals? To address this, we evaluate various pretrained models to assess their suitability

for the AIID task.

4.3.1 Evaluation of pretrained embeddings

The adopted approach to assess whether a model produces suitable embeddings for our task is twofold. First, we evaluate the inherent structure of the embedding space to identify whether natural clusters occur around individual categories, species, and taxa. The idea behind this approach is connected to the idea that a good representation that is useful for hierarchical AIID should generate similar embeddings for examples of the same category and thus natural clusters should emerge. In this study we measure cluster quality through **silhouette score** (Sil) (see details in Section 2.3.4. This is a widely used metric in clustering analysis that expresses both the degree of clusters separation and compactness.

Second, we compute Beecher's information (H_S),(see Section 2.2.2), directly from the pretrained embeddings, to measure the potential of these representations to distinguish between different identities. The H_S calculation requires all 'traits' to be independent, for this purpose we further project the pretrained embeddings onto an orthogonal space through principal component analysis (PCA) (see Section 2.3.4), that optimises variation along the new dimensions. For this H_S computation we make use of the R implementation[2] developed in [Linhart et al., 2019].

This assessment serves as a baseline to establish the quality of the embeddings before applying more advanced algorithms. The goal is to ensure that the embeddings provide a reasonable starting point, capturing sufficient structure to support the work of subsequent methods.

The pretrained models chosen for this comparison are openly available and have been trained on extensive datasets for a variety of audio-related tasks. Their selection reflects the state of the art at the time these preliminary experiments were conducted. Additionally, these models were among the most accessible for embedding extraction, with many sourced from the HEAR Challenge implementation Turian et al. [2022].

CREPE: Released in 2018 , CREPE [Kim et al., 2018], is a deep convolutional neural network (CNN) designed for audio pitch detection. It processes short windows of raw audio directly, avoiding the need for pre-computed features like spectrograms. The model is lightweight and optimized for computational efficiency in pitch-related tasks. The produced embeddings have 2048 dimensions.

[2] `https://cran.r-project.org/web/packages/IDmeasurer/index.html`

VGGish: Introduced in 2017, VGGish [Chen et al., 2020a] is a modified version of the VGG architecture for general audio classification and representation learning. It uses log-mel spectrograms and was pretrained on a large dataset of YouTube audio - Audioset. The model contains approximately 60 million parameters and is commonly used for tasks like sound event detection and music classification. VGGish embeddings have 128 dimensions.

BirdNET: Released in 2021, BirdNET, [Kahl et al., 2021], is a deep learning model for bird species classification. Its architecture combines convolutional and recurrent layers and is designed to handle a large number of bird species. In this work, this is the only model that was trained with data and task specifically in the bioacoustics domain. The embeddings extracted have 1024 dimensions.

OpenL3: Published in 2019, OpenL3 [Cramer et al., 2019], is an audio representation learning model based on the Look, Listen, and Learn (L3) framework. It utilizes CNN backbones such as MobileNet to generate embeddings with 512 dimensions from log-mel spectrograms. here, two versions of the model are considered, **Openl3_music** has been trained on music sounds, and **Openl3_env** which has been trained on environmental sounds.

4.3.2 Evaluation results with Silhouette scores and Beecher information

Silhouette scores for the embeddings of the test set(set of new examples of individuals already used for training), are calculated across the 4 models described above. Results are presented in Figure 4.4. For all hierarchical levels of the problem, Openl3 shows better silhouette scores, and between the two variants of Openl3, there is only a slight advantageous at certain levels. The very low silhouette scores on the ID level reflect the difficulty in finding a representation that (used directly) is able to create distinct and well separated groups for each ID in this multispecies dataset.

Table 4.3 presents the H_S values computed on the PCA transformed embeddings of the test set with various dimensions. These results also support the previous observations with silhouette score. OpenL3_env shows better capacity to encode all the different individuals, particularly at higher dimensions. In general the potential to distinguish individuals is higher as the number of representation dimensions increases. However it is notable the higher potential of the BirdNet model at lower dimensions. Silhouette scores and beecher informa-

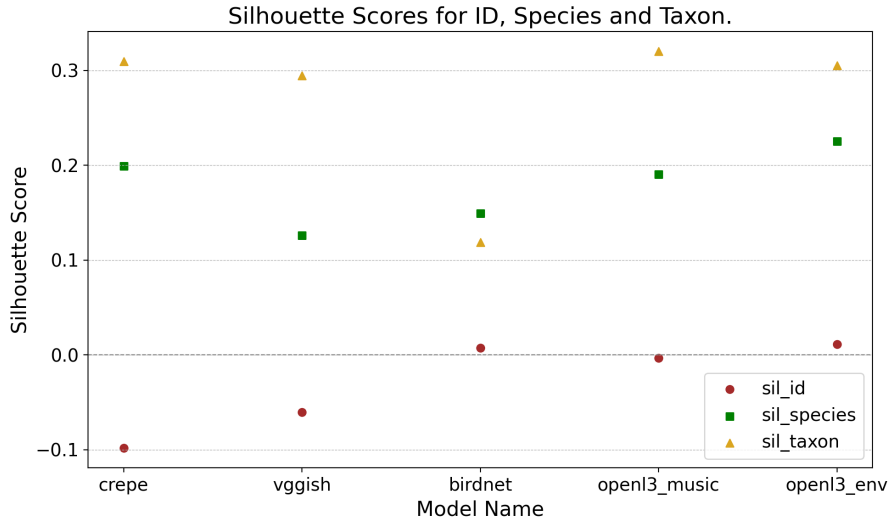


Figure 4.4: Silhouette scores (Sil), are computed on the test set for the 3 levels of the hierarchy considered: ID (in red), species (in green) and taxon (in yellow). Sil scores range between -1 and 1, values closer to 1 are representative of groups that show both better cohesion and separation from other groups.

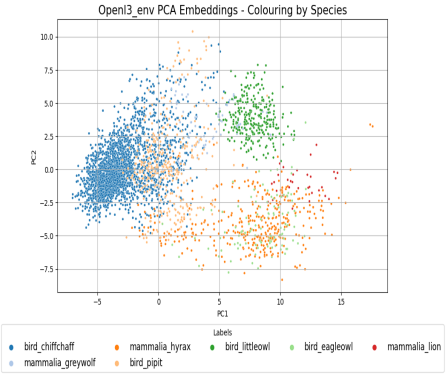
tion provide concrete measures regarding the suitability of these embeddings for the AIID task, however these are meaningless as standalone metrics. Although not exact, visualising the actual embeddings is a great way to conceptualise and gain insights regarding quality of the embeddings. Figure 4.5 provides a visualisation of the spatial distribution of the Openl3_env embeddings of the test set, coloured by the 3 hierarchical categories of the problem (Visualisations produced for the other models can also be found in appendix **??**).

Dimensions	CREPE	VGGish	BirdNet	OpenL3 (Music)	OpenL3 (Env)
3	2.27	2.33	**3.34**	2.23	2.45
12	6.26	5	**7.43**	6.69	7.14
72	15.42	12.72	18.99	19.39	**20.5**

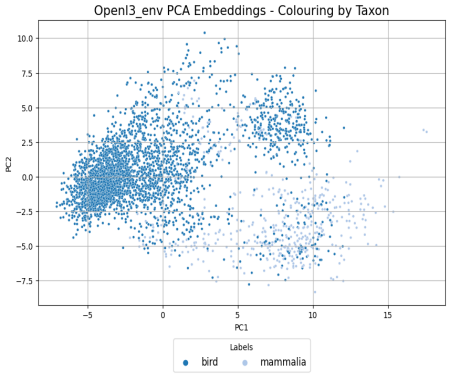
Table 4.3: H_S values for the test set based on various pretrained embeddings. In here we experiment with different number of dimensions when transforming the embeddings through PCA.

4.4 Discussion

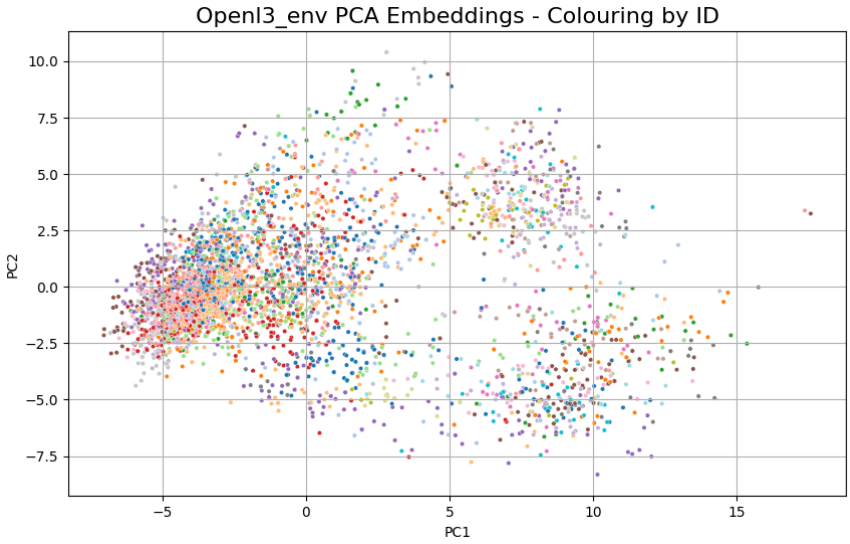
The multispecies dataset compiled for this work provides a rich and diverse foundation for exploring and developing general methods for AIID. The diversity within the dataset encompasses species known to produce acoustic signals (AS),



Colouring by Species



Colouring by Taxon



Colouring by ID

Figure 4.5: OpenL3_env PCA Embeddings visualized using different colouring schemes.

many of which have been extensively studied. For most species, researchers have identified specific acoustic features critical for encoding individuality, offering a valuable reference point for analysis.

Despite the insights gained from traditional feature engineering approaches, the development of generalizable AIID models across multiple species requires a shift toward more universal representations. These representations are not tied to manually selected features but instead leverage advancements in machine learning, where pretrained embeddings have emerged as a promising solution.

In this work, we chose the OpenL3 model to generate embeddings, reflecting the state of machine learning models when this project started. Approximately

three years ago, we conducted pilot studies comparing several pretrained models in terms of their ability to preserve identity-related information. OpenL3 embeddings demonstrated strong performance, showing better separation of calls at multiple levels of the AIID problem (ID, species, and Tax). This result justified the adoption of OpenL3 as the primary embedding generator for our study.

The field has advanced significantly. Currently, more comprehensive and sophisticated models have been released, and, perhaps more importantly, bioacoustic-specific models have become increasingly available [3]. These specialized models are designed with animal vocalizations in mind and are expected to provide representations that are better aligned with the characteristics of bioacoustic data. Comparative studies can help assess whether these models outperform general-purpose embeddings like OpenL3 in terms of preserving identity information across diverse species.

4.4.1 Dataset limitations

Certain aspects related to AIID cannot be fully captured or addressed within the scope of this work. A primary limitation is the variability of AS under different conditions. This encompasses call type dependency, temporal stability, developmental changes, and geographical variations. Different call types within a species may have different AS, as shown in [Elie and Theunissen, 2018] for the case of zebra finches. Vocalisations can change across seasons, over an individual's lifetime, and across different locations. This aspect if further discussed in Section 7.2, Our dataset does not represent these variations, potentially limiting our understanding of how AS varies across different contexts and conditions. Confounding factors in the data pose significant challenges to machine learning performance, particularly in AIID systems. These factors can include sound artefacts from the environment or recording devices that consistently appear alongside the calls of certain individuals in the dataset. While these artefacts may be present in the training data, they are not necessarily representative of real-world conditions. The critical issue arises when machine learning methods inadvertently learn to identify these artefacts instead of the true AS of individuals. This, can severely impact the generalisation capability of the developed methods, limiting their effectiveness in real-world applications. Works like [Stowell et al., 2018], have begun to address this issue in AIID systems, highlighting the importance of mitigating the effects of confounding factors Finally, overlapping vocalisations are particularly relevant in natural settings, where vocalisations from multiple individuals and species often coincide. In general, the

[3]For an updated list, see: https://github.com/bioacoustic-ai/bacpipe

limiting factor to explore these aspects of AIID is related with the lack of annotated data , but also, as discussed in the previous chapter (Chapter 3), a lack of general-purpose methods.

Addressing these limitations should be a future step for creating more robust AIID systems capable of functioning in the real world, and a potential direction in which to build upon the present dataset.

4.4.2 Data availability

At the time of writing, the dataset used in this work has not yet been publicly released. Most of the species data originate from private collections kindly made available for this research. An exception are the recordings of chiffchaffs, tree pipits, and little owls, which are publicly accessible via Stowell et al. [2018].

Chapter 5

Multi-species AIID with Multi-task learning and Hierarchical classification

The previous chapters established the rationale for transitioning from single-species to multi-species AIID systems, primarily from an applications perspective. While the advantages of having such comprehensive tools for animal communication research and conservation efforts are evident, the technical innovations required to develop multi-species AIID systems warrant deeper examination. This chapter explores these technical aspects and demonstrates how they intersect with certain machine learning paradigms, namely multi-task learning. The incorporation of multiple species into the AIID problem, also presents an opportunity to leverage the inherent hierarchical relationships between taxonomic classes and allows exploration of other structured learning approaches like hierarchical classification.

To this end, in Section 5.1, we present and discuss two methodological frameworks: Cross-Species knowledge transfer and hierarchical learning across taxonomic levels. These ideas are realised in the proposed methods described in Section 5.2.1. For the methods incorporating hierarchical concepts, we introduce (in Section 5.3) a specific evaluation metric that quantifies the consistency of model predictions with respect to the underlying taxonomic hierarchy. The analysis of the proposed methods also include an evaluation of the learned embedding spaces, which can further provide insights about their generalisation capabilities.

Section 5.4 presents the results, and this chapter concludes, in Section 5.5, with a critical analysis of these methods' potential application and limitations.

5.1 From multiple-species AIID to multi-task learning

In Chapter 3, we define the characteristics of real-world bioacoustic tasks and examine how these characteristics influence our design of machine learning approaches. The primary challenges stem from the nature of the datasets, which are typically aggregated from multiple sources, resulting in highly heterogeneous and unbalanced class distributions.

Both the few-shot bioacoustic task and the multi-species AIID task share these dataset characteristics, leading to many common challenges. In practice, within the multi-species AIID task, the data corresponding to each species can be considered as coming from a distinct source. Consequently, both tasks seek a unified approach to address specific objectives across diverse data sources.

However, a key distinction lies in the level of granularity associated with each data source. In the few-shot task, each dataset exhibits varying levels of detail in both its acoustic characteristics and its specific objectives. These objectives may range from species recognition to call-type detection or other classification goals. Because the goals differ across data sources, it is difficult to merge them into a single, cohesive task. As a result, in the few-shot scenario, each data source must be treated independently, without drawing explicit relationships between them.

The multispecies AIID task, in contrast, exhibits more structure. All data sources (i.e., species datasets) share the same hierarchical labelling system, comprising three levels: Taxon, Species, and ID. This consistency allows relationships to be established across data sources and leveraging the inherent structure of the dataset to improve learning and generalization.

The ability to establish relationships between tasks is particularly important in scenarios with limited labelled data, such as AIID. By leveraging related tasks, we can mitigate the challenges of small datasets and formulate the problem in a way that enables knowledge transfer across different species and tasks.

The underlying assumption we make for the AIID task is that single species AIID methods, that are trained on "enough" data and allowed to fine tune to the species-specific patterns, should be very successful. However obtaining enough data with ID labels is seldom possible. This makes single species systems hard to scale up as we want to extend AIID to other species, which lead us to explore alternatives to the traditional paradigm. In the centre of our exploration is the idea that by optimising AIID for multiple species together, the machines can learn a general representation of AIID features and become better across different animals and even generalise to completely new species without needing to retraine extensively.

There are two key ideas to explore in this context:

Cross-Species Knowledge Transfer for AIID We hypothesize that the AIID task for each species can be improved by leveraging shared knowledge across species. This idea is inspired by human learning mechanisms. For instance, when a baby learns to interact with the world and grasp objects, they do not specialize in picking up just one specific object. Instead, they develop a general understanding of shapes, materials, and textures, which allows them to grasp a wide variety of objects.

Similarly, in AIID, the strategy to effectively identify individuals in one particular species might follow the same general principles for many other species, and thus learning on multiple species where individual characteristics are encoded helps developing a general understanding of the process. Consequently, models that are trained in this way are less impacted by shortness of data for particular species and are more capable of generalizing and applying this skill to unseen species and individuals. In short, by training on a diverse set of species, the model can develop a broader representation of individual recognition, making it more adaptable to new contexts.

Hierarchical Learning Across Taxonomic Levels The second idea is that taxon classification, species classification, and multi-species AIID are interrelated tasks that can be structured hierarchically based on levels of similarity. In this framework, models can benefit from a structured learning approach where they first solve broader, simpler tasks before progressing towards more complex and fine-grained classifications.

Taxon classification (e.g., distinguishing between taxa like mammal, bird, or amphibian) is a coarse-level task with fewer, well-separated classes. Species classification is a mid-level task that refines taxon classification by distinguishing between species. Multi-species AIID is the most fine-grained task, requiring differentiation at the individual level where we expect inter-class similarity to be highest. By guiding models through this hierarchical progression, we hypothesize that they can leverage prior knowledge from simpler tasks (e.g., taxon-level patterns) to improve their ability to solve more challenging tasks (e.g., individual-level classification). Besides sharing representations across tasks, hierarchical learning can also work by constraining the solution space of each task on the prediction of the parent task in the hierarchy.

Both cross-species knowledge transfer and hierarchical learning have the potential to improve AIID performance by allowing models to develop more gener-

alizable representations and general understandings of the identification process.

5.2 Methods

The ideas described above align well with the aims of **Multi-task learning (MTL)** and **hierarchical classification** introduced earlier in Sections 2.3.2 and 2.3.5, respectively.

MTL is a machine learning paradigm where multiple related tasks are learned simultaneously rather than in isolation. This means that the learning goal is a combination of several objectives and the model learns to prefer solutions that in general satisfy all objectives. In this work, we focus on the neural network implementation of MTL with **hard parameter sharing** [Ruder, 2017]. In this setup, the network includes a common branch, followed by task-specific branches that specialise on individual tasks (Fig. 2.2). The hard parameter sharing configuration is particularly suitable when tasks are strongly correlated and there is evidence that certain feature representations can be useful across all tasks. Sharing parameters reduces the risk of overfitting, especially when each task has limited data and generalisation is the main goal.

Considering hierarchical relationships in the data, besides the sharing of features across hierarchical related tasks, we also adopt concepts from hierarchical classification methods. Namely the hierarchical conditioning of predictions similar to the **Top Down approach** described in Section 2.3.5.

Finally, All the methods developed in this chapter use the **Cross entropy loss function** as the training objective, which is widely used for classification tasks in the closed set scenario, (see Section 2.3).

5.2.1 Proposed methods

This section describes the proposed methods designed to explore the relationships across species and taxonomic levels described above. A summary of these methods is presented in Table 5.1.

Baseline methods: Single task approach.

To establish a baseline, we first develop a flat classification approach, **Multi-species AIID (MS-AIID)**, in which each ID in the training set is treated as an independent class. The network is trained to classify all the 66 IDs without considering taxonomic relationships or joint learning. Additionally, to highlight the challenges associated with the multi-species approach, we implement **Single Species AIID (SiS-AIID)**, where a separate model is trained for each species.

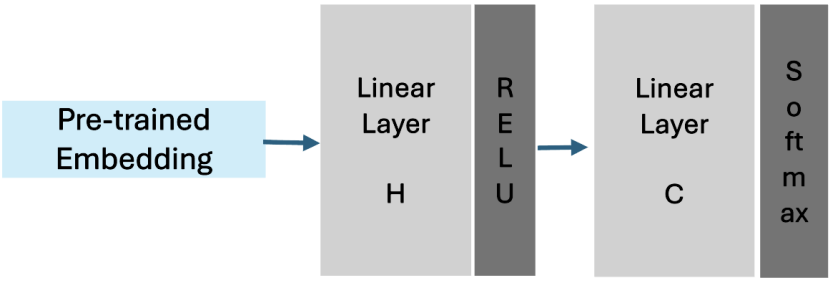


Figure 5.1: Single task network scheme. MS-AIID, SiS-AIID and the Species and taxon classifiers all use this network architecture. The input is the pre-trained embeddings previously extracted from the OpenL3 model (described in chapter 4). These are fed into a single linear layer of dimension H with a RELU on top. The classification layer is the last one of dimension C the same as the number of classes in our task. The classification layer generates logits which are then used to compute the loss function and train the network. In the inference phase, logits are instead transformed into probabilities, through the softmax layer and the class with highest probability is assigned to the input example.

Both methods use the same network architecture (Fig. 5.1). The key difference lies in the number of target classes: **MS-AIID** trains a single model for all species' IDs, while **SiS-AIID** trains distinct models per species. Training also follows the same process for both approaches. Batches of pretrained embeddings are fed into the network, which goes through iterations of gradient descent until early stopping is triggered. Training parameters are summarised in Table 5.1.

Cross-species knowledge transfer methods

To investigate the benefits of feature sharing across species, we develop **Multi-species multi-task AIID (MS-MTL)**, which follows the Multi-Task Learning (MTL) paradigm. The Network architecture follows the schematic presented in Fig. 5.2, where a common branch extracts shared representations across species, and seven task-specific branches are dedicated to optimizing for species-specific AIID task.

In this setup, species-specific AIID tasks are trained on data from their own species only, but all species contribute to training the shared representation. During training, batches are constructed separately for each species and iterated over in a balanced cycle. At each training iteration, logits are computed for the species-specific branches, and their individual losses are combined before back-propagation. This means the shared layers are updated jointly using gradients from all species, while each species-specific classification head is updated only

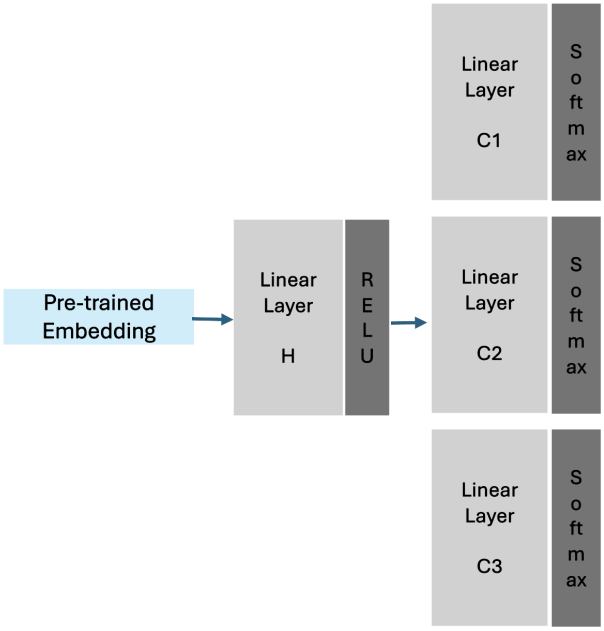


Figure 5.2: Network implementation of Multi-task learning based approaches. Similar to the single task network above, the network contains a single hidden layer of dimension H, this is also the common branch that is shared across all the tasks. This schematic shows 3 task-specific branches, with a single classification layer for each. The dimension of this layer is the number of classes for each task. This setup is used in the MS-MTL and H-MTL approaches. In MS-MTL, there is a branch for each species, each with dimension (C) equal to the number of ID classes for that species. In H-MTL, we define 3 branches for the hierarchical tasks: Taxon classification (C=2), Species classification (C=7) and individual identification (C=66).

using its own loss. This design allows the shared feature extractor to generalize across species, benefiting especially those with limited data.

The batch building strategy adopted is particularly important given the highly unbalanced nature of the dataset. The goal is to ensure that all species contribute equally to the learning of the shared layers and that no single species dominates the process. This should help prevent the network from disproportionately adapting to species with a higher representation in the dataset and forgetting the contributions of the least represented species.

Inference with this model requires knowing the species label of the query example. During this step only one species-specific branch is activated and a single prediction is done. Variations of this implementation where inference is done with all branches and does not require previous knowledge of the species, have been attempted without success. Alternatives and implications of these

are discussed in Section 5.5.

Hierarchical learning based methods

Previously we discussed how hierarchical information can improve AIID classification in two ways:

1. Constraining the solution space by enforcing taxonomic dependencies.
2. Through joint learning and sharing features across taxonomic levels.

To analyse the impact of hierarchical constraints in AIID, we implement **Naïve Hierarchical AIID (NH-AIID)**, For this approach, we first train a taxon and single-taxon species classifiers using the single-task network setup (Fig. 5.1). We also make use of the previously trained single-species AIID classifiers. At inference time predictions at each taxonomic level are conditioned on the predictions from the preceding level. We experiment with a soft hierarchical inference strategy, where probabilities for each individual ID are computed taking into account contributions from all possible taxon and species paths. At each level of the hierarchy, predictions from every relevant model are combined and weighted by their associated probabilities. This allows uncertain predictions at higher levels to still influence the final ID classification, rather than committing to a single path.

The final ID prediction ($\widehat{\text{id}}$) can be defined as:

$$\widehat{\text{id}} = \arg\max_{\text{id}} \sum_{\text{taxon}} \sum_{\text{species} \in \text{taxon}} P(\text{taxon}) \cdot P(\text{species} \mid \text{taxon}) \cdot P(\text{id} \mid \text{species}) \tag{5.1}$$

This approach ensures that lower-level classifications are explicitly guided by higher-level taxonomic predictions, enforcing hierarchical consistency. However, this is inherently rigid, as it requires both prior knowledge of the hierarchical structure and training each single-species AIID classifier as independent models. Consequently, dependencies between species and taxonomic levels are not learned jointly.

To investigate the advantages of shared features across taxonomic levels, we develop **Hierarchical Multi-Task Learning (H-MTL)**. This approach is implemented with the MTL network in Fig.5.2, with 3 branch specific tasks: Taxon classification, species classification and AIID (individual classification). In contrast with the MS-MTL approach, where each species-specific branch is trained only on the corresponding species data, the H-MTL framework allows all data to contribute to all three classification tasks and inference produces a

prediction from every branch. Moreover this approach includes the AIID task in a broader hierarchical multilabel classification task.

The training process follows the multi-task learning paradigm with hard parameter sharing, where a shared feature extractor is optimized jointly for all three classification tasks. First, the pretrained embeddings previously extracted are fed through the network and into three separate classification heads, corresponding to taxon, species, and individual-level identification (AIID).

The total loss is a weighted sum of the three task-specific losses, each computed using cross-entropy:

$$L_{\text{H-MTL}} = \lambda_1 L_{\text{Taxon}} + \lambda_2 L_{\text{Species}} + \lambda_3 L_{\text{ID}} \tag{5.2}$$

where λ are parameters that define the weight given to the contribution of each task towards the total loss.

Finally, in **Hierarchical Multi-Task Learning with Conditioning**, we combine both hierarchical constraints and shared feature learning in a single method. This approach enforces hierarchical consistency while allowing the model to jointly optimize across taxonomic levels.

The model and training process remain mainly the same as in the H-MTL approach, however the prediction at the species and ID levels are conditioned on the outputs from higher levels in the hierarchy. This conditioning is included in the loss function in Equation 5.3, which captures both joint learning aspect and the hierarchical structure imposed.

$$\begin{aligned}
\mathcal{L}_{\text{total}} =& \lambda_{\text{taxon}} \cdot \mathcal{L}\left(\log P(\text{taxon} \mid x), y_{\text{taxon}}\right) \\
&+ \lambda_{\text{species}} \cdot \mathcal{L}\left(\log \tilde{P}(\text{species}), y_{\text{species}}\right) \\
&+ \lambda_{\text{ID}} \cdot \mathcal{L}\left(\log \tilde{P}(\text{id}), y_{\text{id}}\right) \\
\text{where} \quad & \tilde{P}(\text{species}) = \sum_{\text{taxon}} P(\text{taxon}) \cdot P(\text{species} \mid \text{taxon}), \\
& \tilde{P}(\text{id}) = \sum_{\text{species}} \tilde{P}(\text{species}) \cdot P(\text{id} \mid \text{species})
\end{aligned} \tag{5.3}$$

At inference time, we compute all branches and apply the same conditioning process as in NH_AIID approach, obtaining the final ID predictions as defined in Equation 5.1.

Experiment	Learning Rate	Batch Size	H	C	Description
SiS-AIID	0.001	32	varies	varies	Baselines for single-species AIID.
MS-AIID	0.0001	32	256	66	Baseline for multi-species AIID.
MS-MTL	0.0001	varies	256	varies	Explores benefits of cross-species feature sharing.
NH-AIID	-	-	varies	varies	Enforces hierarchical consistency.
H-MTL	0.0001	32	256	2;7;66	MTL across taxon, species, and ID levels.
H-MTL-const	0.0001	32	256	2;7;66	Combines hierarchical constraints with shared learning.

Table 5.1: Summary proposed methods and hyperparameter details. H is the size of the hidden layer, C is the size of the classification layer. When an entry is 'varies', this means that there are one AIID classifier or branch per species, therefore the number of ID classes varies depending of the species.

5.3 Experimental Setup

The methods proposed and described in the previous sections are summarised in Table 5.1, along with their corresponding implementation details and training configurations.

All models are trained using the training and validation splits defined in Table 4.2. Optimisation is performed using stochastic gradient descent (SGD) for weight updates, and an early stopping strategy is adopted, based on the validation accuracy, where training is stopped if no improvement is observed after 20 validation steps, (see Section 2.3).

5.3.1 Evaluation

The methods proposed in this chapter are explicitly designed to address the **closed-set** form of the AIID task. This means that the training goal of the systems is to identify the animal that is vocalizing from a pool of known individuals. Closed-set classification is the most common setup for supervised learning methods, where we evaluate whether the models have effectively learned to classify the exact individuals seen during training.

For this purpose, evaluation is performed on the Test set described in Section 4.2, which consists of several examples of each ID class from the training set. Due to the highly unbalanced nature of the dataset, traditional averaged accuracy metrics may not provide a complete picture of model performance. Models are evaluated instead with a balanced accuracy metric which is an average of recall per class. This allows for a better understanding regarding the benefits of joint learning, particularly for under represented classes.

Balanced accuracy is computed mainly for the ID level, however, for the hierarchical based methods, i.e. methods that also produce predictions on taxon and species levels of the hierarchy, the same metric is computed from predictions at the other levels.

$$\begin{aligned} \text{Balanced Accuracy} &= \frac{1}{C}\sum_{i=1}^{C}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \\ \text{where} \quad & C \text{ is the number of classes,} \\ & \text{TP}_i \text{ is the number of true positives for class } i, \\ & \text{FN}_i \text{ is the number of false negatives for class } i. \end{aligned} \tag{5.4}$$

Beyond standard classification metrics, we introduce a metric designed to quantify errors where predictions violate the hierarchical structure of the taxonomy. **Hierarchical Consistency Errors (Herrors)** counts the number of cases where the predictions at different levels of the label hierarchy are inconsistent with one another.

We define two types of Herrors:

$$\begin{aligned} \text{Herrors}_{\text{species-tax}} &= \sum_{i=1}^{N} \mathbb{1}\!\!\!/\left(H_t(\hat{s}_i) \neq \hat{t}_i\right) \\ \text{Herrors}_{\text{ID-species}} &= \sum_{i=1}^{N} \mathbb{1}\!\!\!/\left(H_s(\hat{c}_i) \neq \hat{s}_i\right) \end{aligned} \tag{5.5}$$

where

$\hat{t}_i, \hat{s}_i, \hat{c}_i$	are the predicted taxon, species, and individual ID class for sample i,
$H_t(\hat{s}_i)$	is the expected taxon for the predicted species $\hat{s}_i$,
$H_s(\hat{c}_i)$	is the expected species for the predicted ID class $\hat{c}_i$,
N	is the total number of samples,
$\mathbb{1}\!\!\!/(\cdot)$	is the indicator function returning 1 if the condition is true, and 0 otherwise.

Evaluation of embedding spaces:

Furthermore, we analyse the embedding spaces learned by the different models to assess their ability to represent both seen and novel classes. This includes data from the standard test set, as well as the Unseen ID and Unseen Species test sets. Evaluating the structure of the embedding space for previously unseen classes offers insight into the generalisation capacity of the models. For this evaluation, we select only the models that learn a multispecies embedding space,

namely MS-AIID, MS-MTL and H-MTL. For each of the three evaluation sets (Test, Unseen ID, and Unseen Species), we extract embeddings from the trained models, from the layer immediately preceding the classification layer. For multi-task learning (MTL) models, embeddings are extracted just after the shared layers, before the task-specific branches. All embeddings are of 256 dimensions.

We evaluated the embeddings using three complementary approaches:

1. **Silhouette Scores:** Silhouette scores are computed at each level of the taxonomic hierarchy (taxon, species, individual), following the method defined Section. 2.3.4. These scores assess how well the embeddings group similar samples and separate different classes. We compute silhouette scores under three conditions: on the test set alone, on the test set combined with the Unseen ID set, and on the test set combined with the Unseen Species set. The test-only scores serve as a baseline, while the other two cases reveal how well novel identities and species are integrated into the embedding space.

2. **UMAP Visualisation:** We apply a two-dimensional UMAP projection (see Section 2.3.4), to the embedding spaces to facilitate visual interpretation of the learned structures. These visualisations support and complement the silhouette scores by revealing clustering patterns and class overlaps. Embeddings are coloured according to the level being analysed (e.g., species or ID) and the evaluation set to which they belong in order to highlight organisational patterns.

3. **kNN Classification:** We evaluate the separability of the learned embeddings using k-nearest neighbours (kNN) classification with $k = 1$ (see Section 2.3.4). The kNN algorithm assigns a class to a query sample based on the ground-truth labels of its nearest neighbours in a reference set. Here, three evaluation setups are considered that differ in the composition of the reference and query data:

   Test: This is the standard kNN classification where the reference is the training set, and the evaluation is done on the test set. This setup reflects how well the embedding space captures the classes used during training. While the results are often aligned with classifier accuracy, they provide a baseline for comparison with the other setups.

   Novel-All: Classification of the Unseen ID and Unseen Species test sets using a reference set composed of the training data and all remaining unseen samples (excluding the query). This setup evaluates the embedding space's ability to organise and represent novel classes not

seen during training. Combined with silhouette scores and visualisation tools, it provides insight into the model's generalisation capabilities.

Novel 1-shot A one-shot classification setup where only a single exemplar per unseen class is included in the reference set. This approximates an open world scenario in which a novel class appears with minimal reference data, and assesses the practical potential of the embedding models for real-world application.

5.4 Results

Accuracy results on the AIID task are averaged per species and across all ID classes. These scores are presented in tables 5.2 and 5.3 respectively. Wherever relevant for the present analysis, targeted comparisons are included here, while the full results can be visualised in Appendix **??**.

Species	SiS-AIID	MS-AIID	MS-MTL
Chiffchaffs	0.94	0.65	0.47
Tree pipits	0.92	0.68	0.89
Little owls	0.96	0.75	0.94
Eurasia eagle owls	0.91	0.78	1.00
Hyrax	0.84	0.50	0.75
African lions	1.00	0.60	1.00
Grey wolves	1.00	0.96	0.96
Overall balanced accuracy ID	0.96	0.66	0.75

Table 5.2: ID Accuracy results per species and overall accuracy for non-hierarchical experimental approaches.

The confusion matrices (CM) for SiS-AIID, MS-AIID and MS-MTL are presented in Fig. 5.3, and the confusion matrices for NH-AIID, H-MTL and H-MTL-const for both ID level and species level are presented in Fig. 5.4. These figures provide a visualisation of the type of errors the models make and particularly which classes are being mixed up in the classification.

Results for the evaluation of the embedding spaces split in 3 parts: 1) Silhouette scores are presented in Table 5.5; 2) Visualisation of MS-AIID and H-MTL-const embedding spaces with UMAP are presented in Figures 5.5 and 5.6, remaining UMAP visualisations are present in the appendix **??**; 3) kNN classification accuracy results are shown in Table 5.4.

Species	NH-AIID	H-MTL	H-MTL-const
Chiffchaffs	0.94	0.93	0.90
Tree pipits	0.83	0.94	0.94
Little owls	0.96	0.96	0.98
Eurasia eagle owls	0.82	0.97	0.97
Hyrax	0.83	0.84	0.88
African lions	1.00	1.00	1.00
Grey wolves	1.00	1.00	1.00
Overall balanced accuracy ID	0.90	0.92	0.93
Overall balanced accuracy Species	0.97	1.00	1.00
Overall balanced accuracy Taxon	0.99	1.00	1.00
Herrors species/taxon	0	0	0
Herrors ID/species	8	6	1

Table 5.3: Results for the hierarchical based methods. This table includes ID balanced accuracy results averaged per species and over whole dataset; overall accuracy scores for species and taxon classification tasks; Also includes Hierarchy consistency errors (Herrors) for both levels considered.

5.5 Discussion

Multispecies AIID approached as single task without without any external information or structured training, is a very challenging task. This is apparent from the large performance gap between **MS-AIID** and **SiS-AIID**, while SiS-AIID can specialize per species, MS-AIID must learn to distinguish individuals across highly heterogeneous classes. As expected, this results in reduced accuracy and increased confusion, both within and across species. The confusion matrices (Figure **??**) show that MS-AIID exhibits inter-species mistakes, which are impossible in SiS-AIID due to isolated species-specific training, and confusion across ID classes within species occurs mainly within chiffchaffs and hyraxes.

Feature sharing across species

MS-MTL introduces joint learning and shared feature representations for AIID across different species. Compared to **MS-AIID**, it performs consistently better for most species, particularly tree pipits and eagle owls. This suggests that features may be transferable across species and beneficial for the AIID task.

However, since **MS-MTL** relies on prior species knowledge at inference time, producing a single prediction only from the corresponding species-specific branch, the performance gain is more likely to be due to the preselection of the specific branch to use, and consequently the decrease of inter-species confusion.

A **MS-MTL** variant in which the final prediction is selected from across all species-specific branches has been attempted. However, performance in this

	kNN accuracy	MS-AIID	MS-MTL	H-MTL	H-MTL-const
Test set	ID	0.90	0.90	**0.92**	0.91
	Species	0.85	**0.99**	**0.99**	**0.99**
	Taxon	0.88	**1.00**	**1.00**	**1.00**
U-id set	ID (Novel ALL)	0.90	0.90	**0.92**	**0.92**
	Species (Novel ALL)	0.99	0.99	**1.00**	**1.00**
	Taxon (Novel ALL)	**1.00**	**1.00**	**1.00**	**1.00**
	ID (1-shot)	0.21	0.19	0.23	**0.25**
	Species (1-shot)	0.94	0.89	0.96	**0.97**
	Taxon (1-shot)	0.98	0.95	**0.99**	**0.99**
U-species set	ID (Novel ALL)	0.29	**0.39**	0.35	0.35
	Species (Novel ALL)	**1.00**	**1.00**	**1.00**	**1.00**
	Taxon (Novel ALL)	**1.00**	**1.00**	**1.00**	**1.00**
	ID (1-shot)	0.93	0.75	0.93	**0.95**
	Species (1-shot)	0.93	0.75	0.93	**0.95**
	Taxon (1-shot)	**0.99**	**0.97**	**0.97**	0.97

Table 5.4: kNN classification accuracy for embeddings produced by four models: MS-AIID, MS-MTL, H-MTL, and H-MTLconst. Evaluated across multiple evaluation sets, results are shown for three levels of classification: individual (ID), species, and taxon. The Test set section includes samples from individuals and species seen during training. The U-id set evaluates generalisation to unseen individuals from known species, while the U-species set tests generalisation to individuals from entirely unseen species. Within each setting, we report accuracy both on the whole set (Novel ALL) and under a more challenging 1-shot condition, where each novel class has only one support example.

setting decreases considerably, primarily due to large discrepancies in prediction confidence across different branches. Branches that were fine-tuned on their data for longer tend to produce more confident predictions, often leading to a bias towards selecting the output of a single branch. Improved regularization strategies across branches, as well as more balanced training across species, may be necessary to mitigate these issues.

Including Hierarchical structure

The methods that include hierarchical information show consistently good performance across all three tasks of the hierarchy.

In **NH-AIID** (Tab. 5.3) we confirm the effectiveness of a naive hierarchical classification approach, which works primarily by constraining the solution space of lower level tasks based on the predictions of the higher-level ones.

This constraint is applied only during inference, meaning that no hierarchical information is used to guide the model during training. Furthermore, this method requires training multiple models in isolation (i.e., separate models for

Dataset	Level	MS-AIID	MS-MTL	H-MTL	H-MTL-const
Test	ID	0.05	0.04	**0.11**	0.10
	Species	0.35	0.22	0.35	**0.36**
	Taxon	**0.38**	0.31	0.36	0.36
Test + U-id	ID	0.14	0.12	**0.16**	**0.16**
	Species	0.25	0.15	0.26	**0.27**
	Taxon	0.17	**0.20**	0.18	0.19
Test + U-species	ID	0.03	0.02	**0.08**	**0.08**
	Species	0.28	0.23	0.35	**0.36**
	Taxon	0.26	0.25	0.30	**0.31**

Table 5.5: Average silhouette scores for embeddings generated by four models: MS-AIID, MS-MTL, H-MTL, and H-MTL-const. Evaluated across three dataset combinations: the standard test set, the test set combined with unseen individuals from known species (Test + U-id), and the test set combined with unseen species (Test + U-species). Scores are computed for each level of the hierarchy: individual (ID), species, and taxon.

AIID per species, species classification, and taxon classification), which poses scalability challenges.

Observing the confusion matrices for this model (Fig. 5.4), it is notable that inter-species confusion still occurs. The soft conditioning approach means that errors at higher levels of the hierarchy can propagate downward. These results, together with the Herrors for ID/Species (table. 5.3), indicate that some mistakes originate from the species classification component.

When comparing this approach with the **MS-MTL** version, which attempts to make predictions from all branches combined, one would expect **NH-AIID** to suffer from similar issues, namely, differing confidence ranges across branches or models trained separately. We hypothesise that **NH-AIID** indeed encounters this problem, however, the hierarchical conditioning imposed in this model appears to mitigate its effects by effectively weighting or "selecting" the correct species-specific model for ID classification.

In **H-MTL** and **H-MTL-const**, the hierarchy is explicitly incorporated in the training process through multi-task learning. The benefit of sharing features across hierarchical tasks is best captured in the results obtained for **H-MTL**, where by jointly training classifiers for taxon, species, and AIID, the model outperforms previous discussed approaches and achieves comparable performance to the **SiS-AIID.**

The **H-MTL-const** variant further improves performance by constraining inference through hierarchical conditioning, similar to the inference process in **NH-AIID**. This addition leads to slight improvements, in particular, fewer

inter-species confusions, which can be observed in the confusion matrices and Herrors ID/Species errors. Indeed the fact that **H-MTL const** makes less number of these type of errors implies that the joint learning across hierarchical tasks is beneficial to all tasks across the hierarchy.

Earlier in this chapter, we hypothesise that in low-data scenarios (see Fig. 4.1 for a visual representation of the dataset distribution), incorporating related tasks into the learning process can improve performance for certain species through transfer learning from other, more data-rich sources. This can be confirmed in the results of **H-MTL**, reflected in the performance differences between **H-MTL** and **MS-AIID**, particularly for species like Hyraxes, which have fewer samples per individual. Notably, Hyraxes and lions show the highest gains in accuracy between the two methods (+0.34 and +0.40, respectively).

However, another factor likely to contribute to the improved performance of **H-MTL** is the more structured and stratified embedding space it learns. If the embedding space is organized around the hierarchical structure of the tasks, classification at the lowest level (AIID) becomes more tractable. UMAP visualisations (Fig.5.5) provide some qualitative support for this, but the effect is more clearly evidenced by the reduced inter-species confusion observed in the confusion matrices for **H-MTL** (Fig. 5.4), as well as by the silhouette scores (Tab. 5.5) and kNN classification results on the test set (Tab. 5.4), which suggest clear separation between classes at all levels of the hierarchy. When comparing **H-MTL** to **MS-MTL**, it becomes evident that although both rely on multi-task learning, the superior performance of **H-MTL** indicates that explicitly encoding hierarchical relationships, rather than simply sharing feature representations, is a key factor in achieving better individual identification.

Evaluation of Embedding Spaces for novel classes

To evaluate the quality of the embedding spaces, we considered three aspects: silhouette scores (Table 5.5), kNN classification accuracy (Table 5.4), and the Visualisation of the UMAP projections. (Fig. 5.5).

Silhouette scores provide a direct measure of how compact and well-separated the clusters are in the embedding space at different taxonomic levels. **MS-AIID** consistently achieves the lowest silhouette scores across all dataset splits, as expected given it is a simpler model. **MS-MTL** shows slight improvements, but both **H-MTL** and **H-MTL-const** achieve the highest silhouette scores overall. These results suggest that incorporating hierarchical structure during training leads to more coherent and better-aligned representations, particularly at the ID and species levels.

KNN classification results serve as indicators of the potential of the embed-

ding spaces to represent novel classes. While all models perform comparably in the standard test set, differences become more pronounced in the unseen ID (U-id) and unseen species (U-species) settings (Table 5.4). **MS-AIID** and **MS-MTL** exhibit significant drops in performance in these conditions. For example, in the "**U-species set** ID (1-shot)" scenario, **MS-MTL** achieves only 0.75 accuracy, whereas **H-MTL-const** reaches 0.95. Similar patterns are observed for species and taxon-level kNN accuracy, supporting the conclusion that embeddings learned with hierarchical supervision are more generalisable and robust to class novelty. These observations can be supported by the visual representation of the embedding spaces through UMAP projections, (Fig. 5.5). These visualisations are consistent: **MS-AIID** appears to create a more mixed and scattered embedding space, particularly at the ID level when comparing with **H-MTL const**.

Taken together, these embedding evaluations provide strong evidence that hierarchical training leads to more structured and discriminative representations, which in turn translate into better generalisation performance. This is particularly evident in scenarios involving novel classes, such as 1-shot ID classification of unseen individuals. While flat and unstructured models like **MS-AIID** fail to generalise effectively in these settings, structured models such as **H-MTL** and **H-MTL-const** demonstrate a clear advantage in both representational quality and performance.

5.6 Conclusion

This chapter evaluated a range of multi-species approaches to perform AIID. We compare flat, multi-task, and hierarchical models across multiple metrics. The results clearly show that simpler approaches, such as **MS-AIID**, which treat the problem as a flat classification task across all individuals, are limited in their performance and ability to generalise. **MS-AIID** results are a good indicator of the challenging nature of the AIID task in a multispecies scenario.

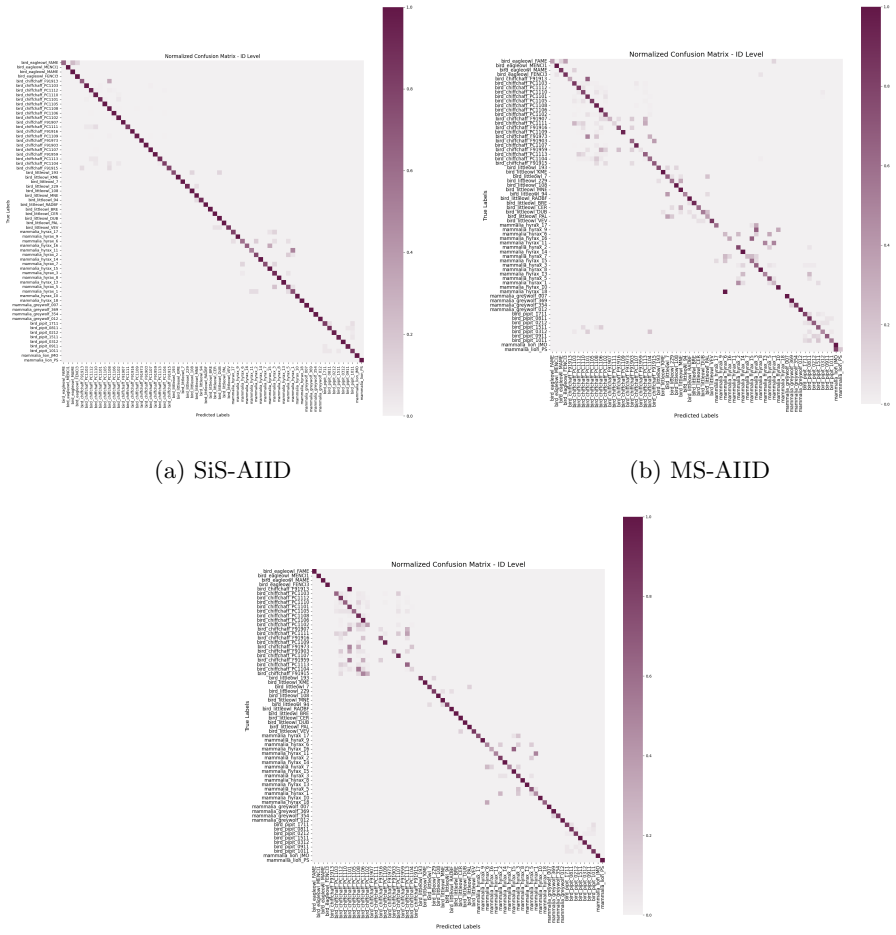
As expected, Species-specific systems (**SiS-AIID**) remain strong performers, especially for well-represented species, but suffer from scalability limitations. **MS-MTL** attempts to bridge this gap through shared representations and species-specific branches, but its performance remains constrained by inference assumptions and imbalanced training.

Hierarchically structured models, particularly **H-MTL** and its constrained variant, achieve the most consistent and robust performance. These models benefit from learning representations jointly across taxonomic levels, leading to improved accuracy, reduced confusion, and better embedding quality. This is supported by higher silhouette scores, stronger KNN accuracy in novel-class

scenarios, and cleaner confusion matrices with fewer errors.

The observed differences in silhouette scores and KNN classification accuracy across models suggest that representation quality is tightly coupled with how well the training process enforces structure in the embedding space. In particular, models that incorporate hierarchy in the training (e.g., **H-MTL** and **H-MTL const**) appear to learn embeddings that are more separable and create better class boundaries at all the taxonomic levels. This is especially relevant when working towards open world classification. These experiments lead us to conclude that indeed imposing some structure in the embedding space can be beneficial when broadening the classification space towards novel classes, i.e. if we aim to develop AIID systems that generalise beyond the closed-set scenario, then optimising models explicitly for the learning of structured, discriminative, and hierarchically-aware embedding spaces becomes necessary.

The next chapter builds directly on this insight by shifting focus to distance based learning approaches, which are specifically designed to optimise the learning of the embedding spaces that support AIID and generalisation to unseen individuals and species.

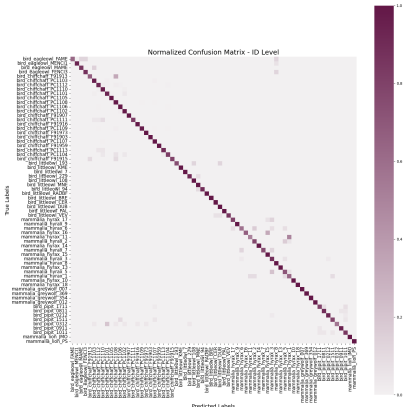


(a) SiS-AIID

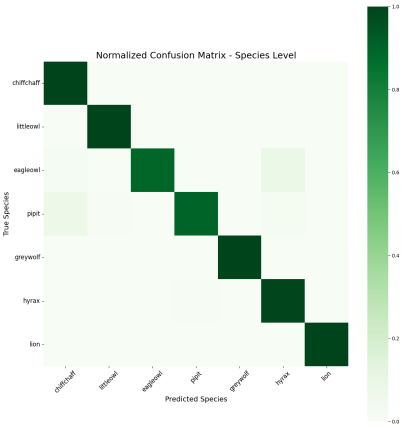
(b) MS-AIID

(c) MS-MTL (strict inference)

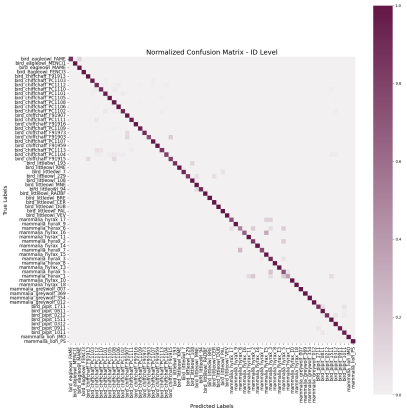
Figure 5.3: Confusion matrices for ID classification predictions on the test set for the models: (a) SiS-AIID, predictions of Id for each species are merged to create this confusion matrix, inter-species confusion is not possible with these models; (b) MS-AIID and (c) MS-MTL.



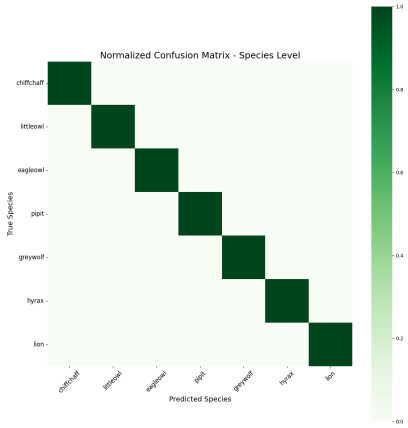
(a) NH-AIID ID level



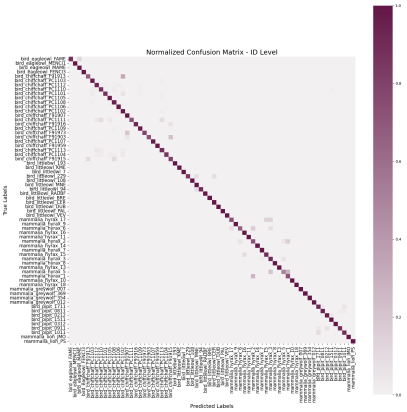
(b) NH-AIID Species level



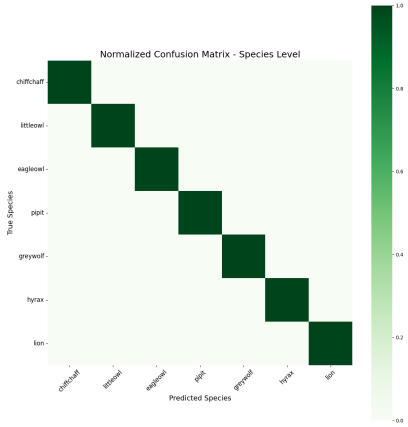
(c) H-MTL ID level



(d) H-MTL species level

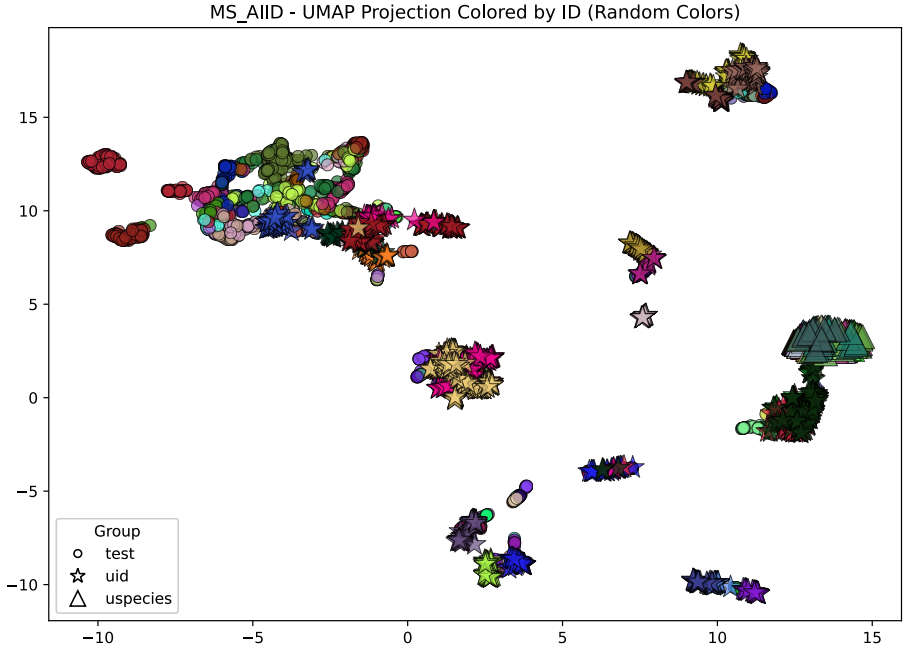


(e) H-MTL-const ID level

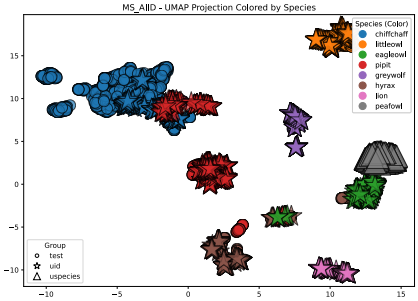


(f) H-MTL-const Species level

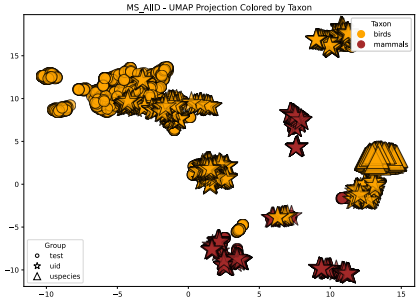
Figure 5.4: Confusion matrices for individual (ID) and species-level predictions generated by the three proposed hierarchical models. (a–b) correspond to NH-AIID, (c–d) to H-MTL, and (e–f) to H-MTL const. Each pair of matrices illustrates model performance at both levels, highlighting inter-species confusion and the effects of hierarchical conditioning.



(a) MS-AIID UMAP ID level

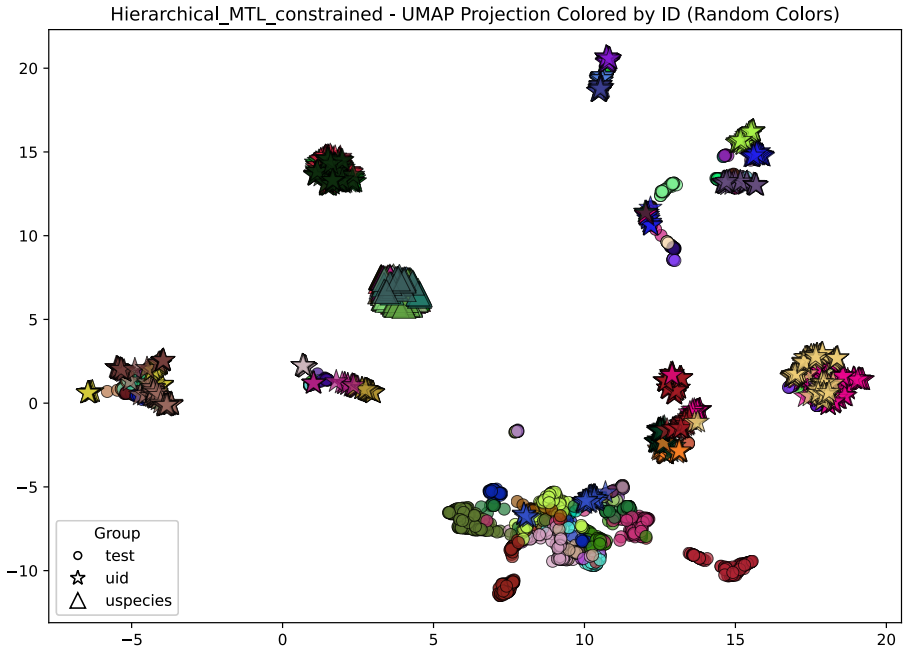


(b) MS-AIID Species level

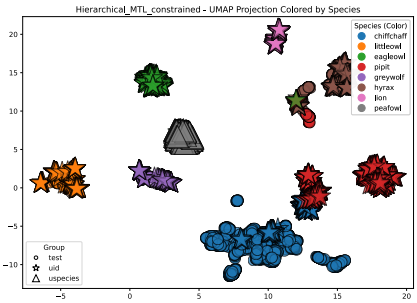


(c) MS-AIID Taxon level

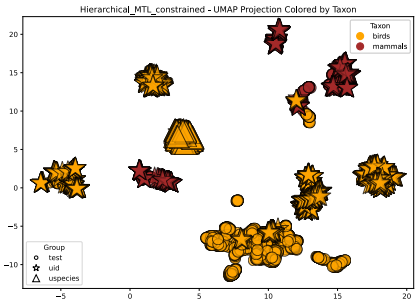
Figure 5.5: UMAP visualisation of the embedding spaces produced by the **MS-AIID** models across all evaluation sets. Each row corresponds to a different hierarchical level: individual (ID), species, and taxon. Colours indicate class membership at each respective level, while marker shapes denote the evaluation subset (test, unseen individuals, or unseen species). These plots highlight differences in embedding structure and class separability.



(a) H-MTL-const ID level



(b) H-MTL-const Species level



(c) H-MTL-const Taxon level

Figure 5.6: UMAP visualisation of the embedding spaces produced by the **H-MTLconst** models across all evaluation sets. The top panel shows the individual (ID) level; the bottom panels show species and taxon levels. Colours indicate class membership at each respective level, while marker shapes denote the evaluation subset (test, unseen individuals, or unseen species). These plots highlight differences in embedding structure and class separability.

Chapter 6

Distance based learning of hierarchical embedding spaces for AIID

AIID systems that can function effectively in real-world scenarios has been our guiding goal throughout this work. Such systems must satisfy two critical requirements: (1) perform AIID across multiple species and (2) be scalable models capable of generalizing to previously unseen classes, allowing for the application to other scenarios where the model was not trained on, thus operating within an open world paradigm. In this chapter we address the second requirement. More specifically we focus on learning embedding spaces that allow or facilitate classification on previously unseen classes.

Open world classification and generalisation to unseen classes can be approached through multiple strategies, not all of which rely on embedding spaces (see Section 2.3.7). However, approaches that learn distance based representations and support geometric reasoning have shown strong performance across many domains, [Mahdavi and Carvalho, 2021]. These methods inherently produce structured metric spaces (embedding representations) that can, in principle, accommodate novel classes/clusters by preserving meaningful relationships among data.

Indeed, throughout this thesis, the analysis of the embedding spaces created has been used primarily as an evaluation tool for deeper understanding of classification outcomes and to gain insights into the generalisation capabilities of the developed methods. However, we have yet to look into the process of learning these embedding spaces. In this chapter, we directly address this by focusing on optimising the embedding learning process and guide it towards our main

objectives. In summary, we shift the objective from building a classification multispecies AIID system to learning embedding spaces whose representation can extend to novel classes and consequently their classification.

Chapter 5 established a direct relationship between the "structure quality" of the embedding space, in particular the presence of hierarchical structure, and their effectiveness in multispecies AIID and generalisation to novel classes. However, that chapter also highlighted limitations associated with the cross-entropy approach, especially concerning novel classes. These findings lead us to move beyond traditional cross-entropy-based methods and adopt a Distance-Based Learning (DBL) paradigm to optimise the learning of embedding spaces that inherently reflect hierarchical structures.

The link between open world classification and DBL methods is further explored in Section 6.1. Subsequently, in Section 6.2, we propose two DBL methodologies for AIID, designed to learn hierarchically structured embedding spaces. Evaluation of the embedding spaces (in Section 6.3.1), follows a similar process as previous chapters, designed to highlight the quality of the hierarchical structure and ability to generalise to unseen classes. The results and discussion of this evaluation are presented in Sections 6.4 and 6.5.

The two DBL methods proposed here have been previously published as conference papers in:

- Nolasco I, Moummad I, Stowell D, Benetos E. **Acoustic identification of individual animals with hierarchical contrastive learning**. In International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2025 Apr 6 (pp. 1-5). IEEE.[Nolasco et al., 2025]

- Nolasco I, Stowell D. **Rank-based loss for learning hierarchical representations.** In International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022 May 23 (pp. 3623-3627). IEEE.[Nolasco and Stowell, 2022]

6.1 Learning embedding spaces with distance-based learning methods

Embedding spaces are naturally formed during the training of neural networks, in that a typical 'hidden' layer outputs a vector of continuous values, and this vector (embedding) can be interpreted as a coordinate in a feature space. Even when the structural organisation of these spaces is not the explicit objective of the training, the emergent structure can be semantically meaningful. In models trained with cross-entropy loss, for example, the embedding space is

shaped primarily to support classification by maximising separation between data points of different classes. A network trained to distinguish between species vocalisations, will typically produce embeddings where each species forms a distinct cluster, allowing a classifier to define clear decision boundaries. In this way, the loss function influences the geometry of the embedding space to serve the specific goals of the task.

Despite this task-specific "shaping", embedding spaces have been shown to capture semantic structure beyond the training objective [Scholl, 2024], which can enhance interpretability and support reasoning beyond the original task. To continue with the species classification example, embeddings may reflect acoustic similarities across species, with species that sound more alike positioned closer together in the space. This organisation suggests the emergence of a semantic structure, even though this structure is not explicitly the goal or guaranteed by the training process. Consequently, any semantic relationships captured in these spaces remain limited.

To address these limitations, DBL offers an alternative approach that explicitly optimises the geometric relationships between data points in the embedding space, (See further in Section. 2.3.3). Unlike cross-entropy training, which is focused on class prediction, DBL methods shape the space by directly modelling similarity and dissimilarity between samples. The fundamental principle is to bring embeddings of samples from the same class closer together while pushing apart those from different classes. By enforcing these relational constraints, DBL creates embedding spaces that are more structured and semantically meaningful.

DBL trained embeddings are particularly well suited for open world classification, where models must be able to reason about novel categories. The rationale behind this is that a well-optimised embedding space should extend its semantic organisation beyond the training categories, allowing meaningful structure to emerge even in regions not populated by known classes. In other words, a semantically structured geometric space enables novel samples to be interpreted and potentially classified based on their relative position to known classes.

Two widely adopted approaches in distance-based learning (DBL) are **Supervised Contrastive Learning** (SupCon) and the **Triplet Loss** (see detailed definitions in Section 2.3.3). Both methods follow the same core principle described above: they aim to minimise distances between examples of the same class while increasing distances between examples of different classes.

The key distinction lies in how comparisons are structured. SupCon uses a *1-vs-all* approach, where for each anchor example, all other examples of the same class in the batch are treated as positives, and all examples from different classes

are treated as negatives. Instead, Triplet loss explicitly compares three samples at a time: an anchor, a positive (same class), and a negative (different class), enforcing a relative margin between the anchor–positive and anchor–negative distances.

These two approaches are the basis of the proposed methods presented next, in which both are adapted and expanded to incorporate hierarchical structure in the embedding space.

6.2 Proposed approach: Learning a hierarchical embedding space

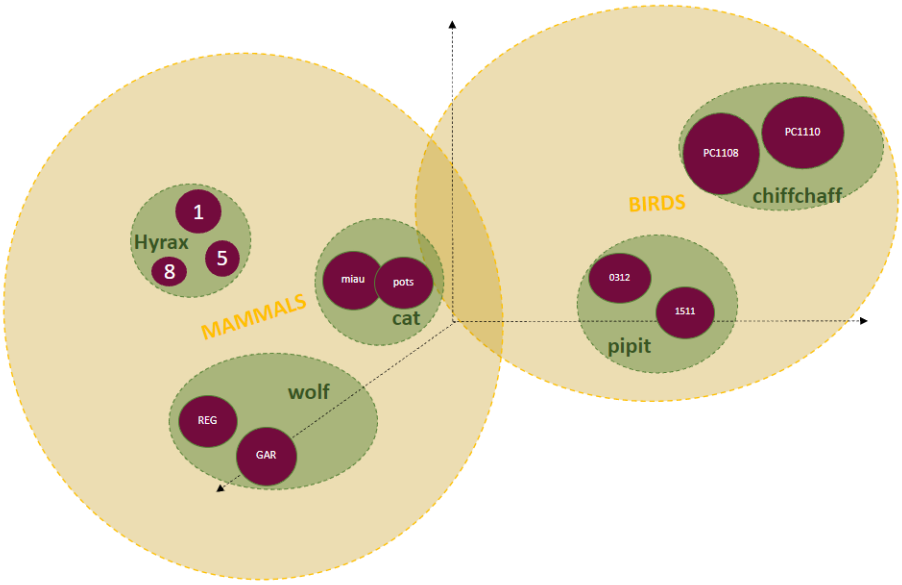


Figure 6.1: Illustration of a structured embedding space that aligns with animal taxonomy hierarchy. In this space, distances between data points correspond to their hierarchical relationships: the yellow blobs represent clusters of different taxa (mammals and birds), these incorporate other clusters representing the species, in green, and within each species cluster, other clusters are formed that represent the various identities, in dark red. As we move down the taxonomic hierarchy, the resulting clusters become progressively more compact and specific.

.

In this chapter we aim to learn embedding spaces that preserve and represent the hierarchical relationships between data points. The desired outcome, a hierarchically structured embedding space, is illustrated in Figure 6.1. This idea is directly influenced from our findings described in Chapter 3, particularly the potential to represent various and distinct categories in a single embedding

space. The hierarchical embedding space proposed here follows that, with the added objective of learning a meaningful hierarchical structure.

An important goal, which further justifies the inclusion of hierarchical structure in the embedding space, is the ability to make incomplete predictions about a vocalising animal. By incomplete we refer to situations where the system is not able to identify the specific individual, perhaps because it has never encountered that animal before. However, it may still be able to determine the broader taxonomic group to which the individual belongs or even the species. In such cases, rather than having the system failing, it should indicate similarity to known classes, even if the closest classes are at other levels such as species or taxonomic groups.

Following the Distance-based learning paradigm we define two loss functions that incorporate this hierarchical structure as the training goal.

6.2.1 Hierarchical Contrastive losses (HCL)

The first approach is based on contrastive learning applied to a hierarchical multi-task (H-MTL) setup. It is a natural follow up from the method developed in the previous chapter (see Section 5.2.1), where we mainly modify the training objective from a cross entropy function to a contrastive loss function. The Hierarchical contrastive loss (HCL) adopted here was originally proposed in Zhang et al. [2022a], where the authors demonstrate its advantages in capturing and preserving hierarchical relationships between examples across various datasets in the image domain. By framing a hierarchical classification problem as multi-label, the authors demonstrate that there are important extra information to be leveraged in the other levels of the hierarchy. This finding aligns with our own findings from the previous chapter.

The basis for HCL is the flat supervised contrastive loss (**SupCon**), introduced in Section. 2.3.3, that is applied to each level of the hierarchy independently. HCL consists on a weighted combination of SupCon loss applied to each level of the hierarchy, similar to how the H-MTL loss function is composed. We adapt three versions of HCL to our problem:

Hierarchical Multi-label Contrastive loss (HiMulCon) is the contrastive loss version of H-MTL (Sec. 5.2.1)). As a note, this loss could perhaps be more accurately named "Multi-task" instead of "Multi-label" given it can be used with examples where each have labels for 3 different tasks, however in order to keep consistency with the work that first proposed this loss [Zhang et al., 2022a] we choose to preserve the multi-label name here.

HiMulCon is formulated as:

$$L_{\text{HiMulCon}} = \sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-\lambda_l}{|P_l(i)|} \sum_{p \in P_l(i)} L_{\text{pair}}(i, p_i^l) \tag{6.1}$$

where L represents different levels in the hierarchy, λ_l is a level-dependent penalty factor, $P_l(i)$ is the set of positive pairs at level l, and L_{pair} calculates the contrastive loss for a specific pair.

Hierarchical Multi-label Contrastive Enforced loss (HiMulConE) In the previous version of the loss, hierarchical relationships between tasks are only implicitly captured through the combination of task-specific losses and the joint optimization objective across levels. However, each task-specific loss component is computed independently of the others, which means that the hierarchy of the labels is not explicitly enforced and hierarchical consistency may not be preserved. To address this limitation, an additional constraint is introduced that defines a minimum value each loss at a given hierarchical level must satisfy. This value is determined by the loss at previous levels in the hierarchy, enforcing the condition that the loss at a parent level must always be lower than or equal to the loss at its corresponding child level. In essence, this encodes the intuition that confidence at more specific levels (e.g., individual identity) should not exceed confidence at more general levels (e.g., species or taxon). For example, if we are confident that the individual is Lion A, we must be equally or more confident that it is a lion, and even more so that it is a mammal.

If we define the highest contrastive loss among all positive sample pairs for level l as:

$$L_{\text{max}}^{\text{pair}}(l) = \max_{(i, p_l^i)} L^{\text{pair}}(i, p_l^i) \tag{6.2}$$

then the combined loss function **HiMulConE** is defined as:

$$L_{\text{HiMulConE}} = \sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-\lambda_l}{|P(i)|} \sum_{p \in P_l(i)} \max(L_{\text{pair}}(i, p_i^l), L_{\text{max}}^{\text{pair}}(l-1)) \tag{6.3}$$

Here, the term $L_{\text{max}}^{\text{pair}}(l-1)$ represents the maximum loss computed from the previous level, and it defines the lower bound for the loss that can be attained at level l. **HiMulConE** combines two types of penalties: an independent level-specific penalty, modulated by the weighting factor λ_l, and a hierarchical consistency constraint that acts as a dependent penalty, based on the losses computed at higher levels in the hierarchy.

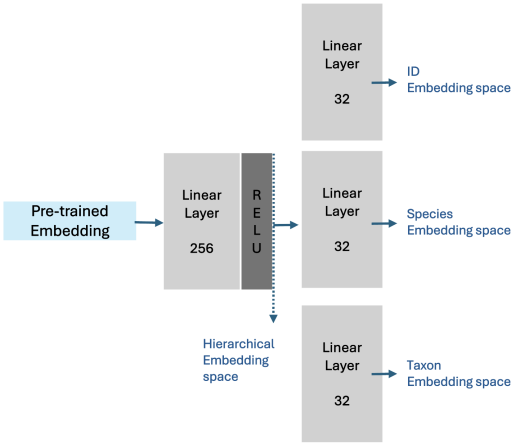


Figure 6.2: Schematic of network trained with the proposed loss functions. The input is the pretrained embeddings previously extracted from the OpenL3 model (described in chapter 4). These are fed into a single linear layer of dimension 256 followed by a RELU. This network is designed to be very similar to the network used with the MTL approaches in the previous chapter, in order to preserve comparability. here instead of a classification layer of dimension C (the same as the number of classes in our tasks) the last layer is replaced by a 32 linear layer.

The training process with these losses follows a similar process as the one used for MTL approaches. However a key difference lies on the validation step that is used to probe the quality of the training process. Instead of computing a validation accuracy from the resulting probability distributions of the softmax, here we apply a kNN algorithm (see Section 2.3.4), with $K = 1$, and classify each example of the validation set against the embedded examples of the training set. Further training details and hyperparameters are described in Table 6.1.

As illustrated in the network scheme (in Figure 6.2), there are several points where embedding spaces can be extracted. First, the branch-specific embedding spaces, namely the ID, species, and taxon embedding spaces shown in the figure, are generated at the end of each branch head. These are optimized independently for their respective objectives and are only expected to retain representations relevant to the other tasks if there is a strong natural overlap in the features that are important across levels. Therefore, while we do not expect these spaces to express the ideal embedding space conceptualised in Figure 6.1, the degree to which they approximate it can further support our claim that acoustic characteristics of animal vocalisations can be organised following the animal taxonomy, as discussed in Section 4.1.

While neither the **HiMulCon** nor the **HiMulConE** approach explicitly aims to align the hierarchical levels within a single embedding space, for ex-

ample, by enforcing that individuals from the same species be closer to each other than to individuals from different species, the multi-task learning (MTL) setup is expected to promote the emergence of such structure. Since the shared layers are supervised simultaneously by all three tasks, the resulting shared embedding space is encouraged to support separability across classes at all levels of the hierarchy. Prior experiments with Hierarchical Multi-Task Learning (HMTL) (see Sections 5.2.1 and 5.5), support the hypothesis that the shared embedding space does reflect the structure in Figure 6.1. Nevertheless, this remains a hypothesis that can only be validated empirically, since the structure is not guaranteed by the loss formulation alone (even in the hierarchical enforced version **HiMulConE**).

To address this conceptual limitation, we introduce, in the next section, a loss function developed to directly incorporate spatial separability across all levels of the hierarchy.

6.2.2 Rank based Loss (RBL)

We first introduced RBL in Nolasco and Stowell [2022], as a method to explicitly train models to produce hierarchically structured embedding spaces, such as the idealised structure in Figure 6.1. RBL can be viewed as a generalisation of the hierarchy-aware loss proposed by Jati et al. [2019], which extends the standard triplet loss to capture hierarchical relationships between examples in a two-level class hierarchy. In their work, a *quadruplet loss* function is proposed, where: (i) the positive example belongs to the same fine-level class as the anchor, (ii) the negative belongs to a different fine-level class but the same coarse-level class, and (iii) the double-negative belongs to a different coarse-level class. Applied to the AIID task, this would translate to: the anchor and positive being from the same individual (ID), the negative from a different individual but the same species, and the double-negative from a different species. However, this formulation cannot naturally incorporate higher levels of hierarchy (e.g., taxon).

RBL was designed to be more flexible and scalable, removing several of the limitations imposed by the quadruplet loss. First, RBL discards the notion of an anchor. Instead of sampling structured quadruplets, it considers all possible pairwise comparisons within a batch. Each pair is assigned a hierarchical relationship, named a rank, which is based on their distance in the class hierarchy (e.g., same ID, same species, different species). This allows the model to learn from an arbitrary number of hierarchical relationships and scale to represent more more complex structures, such as the three-level hierarchy (ID–Species–Taxon) used in our work.

Second, unlike the quadruplet loss—which requires manually specifying mar-

gin values to separate hierarchical levels—RBL avoids pre-defining fixed separation distances between ranks. Instead, it learns target distances directly from the data by observing the distribution of pairwise relationships. This data-driven approach is appealing because it allows natural properties of the signals—such as inter-individual variability or species-specific vocal patterns—to shape the final structure of the embedding space in a more biologically meaningful way. However, this flexibility comes with potential drawbacks. Since the target distances are learned from the initial embedding distribution, the resulting loss can be sensitive to the quality of the initialisation. Additionally, imbalances in the dataset—such as uneven representation of individuals or species—may distort the inferred distance relationships and bias the learning process. As such, while RBL enables adaptive and scalable learning of hierarchical embeddings, it also requires careful consideration of data quality and initial feature representations.

Formally the rank based loss is defined as:

$$L = \frac{1}{P}\sum_{p}^{P}(1 - I_p) * (EmbDist_p - TargetDist_p)^2 \tag{6.4}$$

where $EmbDist_p$ is the distance in the embedding space between two embeddings, $TargetDist_p$ is the target distance for that pair given the rank of the pair, and I_p is a Boolean indicating if the pair is correctly distanced given their rank in the label tree.

Ranks are assigned based on the hierarchical label distance between two examples. Assuming that labels are organized as a tree-structured hierarchy (e.g., taxon, species, individual), the distance between two examples can be defined by the number of nodes that separate them in the tree. This tree-based label distance serves as the basis for assigning discrete ranks between pairs of examples. Specifically: R0 – both examples belong to the same individual; R2 – same species but different individuals; R4 – different species within the same taxonomic group; R6 – different taxonomic group.

While this formulation uses four discrete ranks, the approach is not inherently limited to a fixed number. In principle, RBL can scale to more larger or more granular label structures, as long as the underlying label hierarchy remains well-defined.

The loss is computed in 5 steps described below and a toy example of the process is presented in Figure 6.3:

1. Compute a rank map from the tree of ground truth labels: Each pair of examples has a rank given by the tree distance of their labels.

2. Compute all the pairwise distances in the batch in the embedding space,

and sort them.

3. For each rank, assign a target distance by selecting whatever distance in the sorted distances vector falls at each rank.

4. Compute I_p as: 1 if distance of the pair is within the correct positions in the sorted distances vector, else 0 if distance of the pair is wrong given the ground truth rank.

5. Compute the loss from equation (6.4).

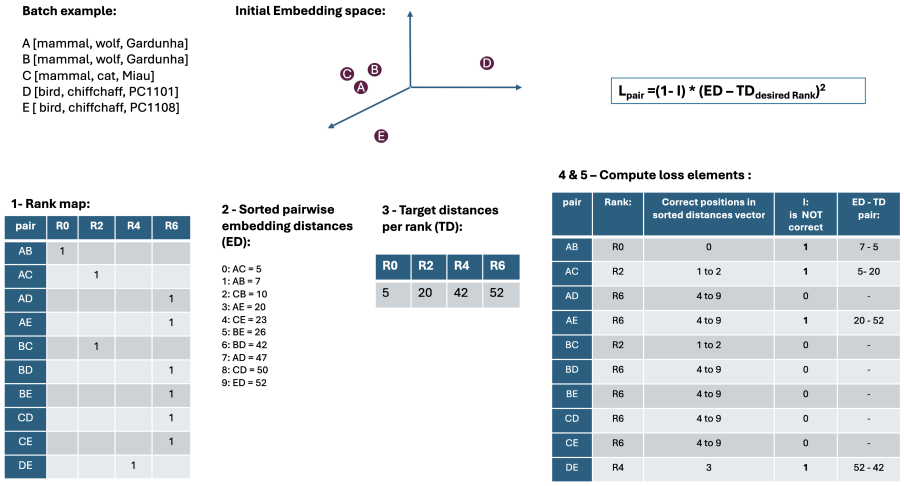


Figure 6.3: Example of RBL computation for one batch of data.

Training with RBL follows a similar process to that used with HCL. However, RBL is significantly more sensitive to the composition of training batches. This is because the loss relies on capturing in the pairwise comparisons all the ranks of the hierarchical structure. To ensure consistent representation from all ranks, we employ a custom batch sampler that constructs batches with a fixed number of samples (e.g 64), but enforces the rule that each batch must contain at least one pair for each of the four ranks (R0, R2, R4, R6). This is achieved by randomly selecting candidate batches and and discarding the ones that do not conform to the rule. While this may imply that not all training samples are used in a given epoch, it guarantees that the loss function sees enough examples from each rank.

6.3 Experimental setup

In this section we describe the experiments conducted with the DBL methods described in Sections 6.2.1 and 6.2.2. Experiments are summarised in Table 6.1

together with their main hyperparameters and characteristics.

The training, validation and evaluation are supported by the same dataset splits (training, Validation and evaluation sets) defined in early chapters (see Section 3.2.1).

Experiment	Learning Rate	Batch Size	Hidden. Dim.	Output Dim.	Description
SupCon	0.001	128	256	32	Non-hierarchical supervised contrastive loss. Only ID-level loss active.
HC	0.001	128	256	32	Trained Hierarchical Multi-label Contrastive loss (Eq. 6.1), with equal task weights.
HCλ	0.001	128	256	32	Tuned hyperparameters for HiMulCon loss (Eq. 6.1). Task weights (ID: 0.5, taxon/species: 0.25).
HCE	0.001	128	256	32	Trained with the Hierarchical Multi-label contrastive Enforced loss (Eq. 6.3), with equal task weights.
HCEλ	0.001	128	256	32	Trained with the Hierarchical Multi-label contrastive Enforced loss (Eq. 6.3), and best parameters as above.
RBL	0.0001	64	*	256	Network trained with Rank based loss (Eq. 6.4)

Table 6.1: Summary of Experiments and their main hyperparameter settings. The Contrastive based approaches generate a unified embedding in the hidden layer. RBL does not train a shared layer, so the embedding is taken from the last layer. Input size for all experiments is 512 from the OpenL3 embeddings.

6.3.1 Evaluation

In this chapter the evaluation process focuses on the assessment of the learnt embedding spaces and how well these encode hierarchical structure into the spatial arrangement

For the HCL-based experiments, the embedding space under evaluation is extracted from the *shared* layers of the network shown in Figure 6.2. This layer outputs a 256-dimensional embedding space, allowing for direct comparison with previous experiments, such as H-MTL, described in Section 5.2.1.

For the RBL experiments, we use the architecture shown in Figure 6.4. In this case, the embedding space is taken from the final layer of the network, which is also configured to produce 256-dimensional embeddings to maintain consistency across evaluations.

This evaluation is conducted in a similar way as to the embedding space

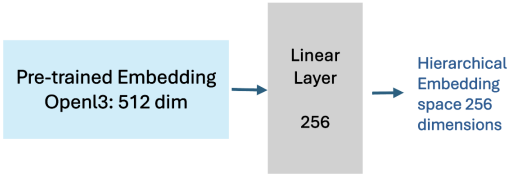


Figure 6.4: Schematic of network used in RBL experiments. The Network consists in a single linear layer that transforms pre-trained embeddings from Openl3 model with 512 dimensions into a 256 dimension embedding.

evaluation performed in the previous chapter (see Section 5.3). It mainly addresses two aspects: 1) Quality of the clusters formed at the various levels of the hierarchy and 2) Suitability of embedding spaces to serve as basis for simple classifiers. The first aspect is captured through the calculation of silhouette scores (see Section. 2.3.4) for the three levels of the hierarchy and for 3 combinations of the evaluation sets: on the test set alone, on the test set combined with the Unseen ID set, and on the test set combined with the Unseen Species set. The test-only scores serve as a baseline, while the other two cases reveal how well novel identities and species are integrated into the embedding space. Regarding the second aspect, this is assessed by applying KNN classifiers (with $k = 1$) to the embeddings extracted from the different models. Similar to the previous chapter (Section 5.3), we apply KNN under four evaluation setups: *Test*, *Novel-ALL*, *Novel 1-shot*, and *Novel 5-shot*. The final *Novel 5-shot* setup was included to provide more complete and nuanced assessment. We consider the few-shot setups to be an approximation to the challenging scenarios encountered in the real world, and thus can provide insights on how the models generalise to novel classes.

In our earlier analysis of pretrained models for data representation (Section 4.3), we observed that some pretrained embeddings, such as those from OpenL3, already exhibit a degree of hierarchical consistency. This observation leads us to question the extent to which our proposed methods actually reshape these initial embedding spaces. To enable a direct comparison between the structural properties of the learned embeddings and the original OpenL3 representations (which serve as the foundation for all experiments in this chapter), we apply the same evaluation process to the initial Openl3 embeddings.

Additionally, we examine the impact of embedding space dimensionality on our evaluation outcomes. To ensure a fair comparison, we apply PCA (see Section 2.3.4), to reduce the dimensionality of the OpenL3 embeddings. For our proposed approaches, we also conduct an ablation study by varying the

number of dimensions in the network's hidden layer – i.e., the layer responsible for generating the embedding space.

Finally this evaluation is also supplemented with UMAP visualisations of the embeddings spaces.

6.4 Results

Silhouette scores and KNN accuracy results on all experiments described above, are presented in Tables 6.2 and 6.3, respectively.

Tables 6.4 and 6.5 present the results from the ablation study on the **HC** experiment and the initial Openl3 embeddings. Dimensions considered are 128, 64, and 32. Further results on this study can be found in Appendix **??**. To aid this analysis, UMAP visualisations of generated embeddings through the Openl3 model and the HCEλ model can be found in Figures 6.6 and 6.5. Additional visualisations with other models are presented in appendix **??**.

To facilitate comparison across learning paradigms- DBL and cross entropy based learning- kNN accuracy and silhouette scores for **HCE** and **H-MTL** approaches (see Sec. 5.2.1) are presented side by side in Tables 6.7 and 6.6.

Dataset	**Level**	**openl3 _pca256**	**SupCon**	**HC**	**HCλ**	**HCE**	**HCEλ**	**RBL**
Test	ID	0.020	0.136	0.183	0.129	**0.196**	0.121	0.049
	Species	0.235	0.320	0.386	0.335	**0.394**	0.320	0.176
	Taxon	0.311	0.288	0.356	0.322	**0.363**	0.312	0.197
	AVERAGE	0.189	0.247	0.309	0.262	**0.318**	0.251	0.141
Test + U-id	ID	0.027	0.093	0.125	0.096	**0.130**	0.092	0.086
	Species	0.194	0.265	0.334	0.290	**0.338**	0.269	0.055
	Taxon	0.226	0.185	0.238	0.222	**0.250**	0.220	0.098
	AVERAGE	0.149	0.181	0.232	0.203	**0.239**	0.194	0.103
Test + U-species	ID	0.014	0.109	0.149	0.106	**0.159**	0.098	0.044
	Species	0.247	0.287	0.362	0.331	**0.374**	0.323	0.117
	Taxon	0.248	0.253	**0.314**	0.279	0.311	0.257	0.146
	AVERAGE	0.17	0.216	0.275	0.239	**0.281**	0.226	0.118

Table 6.2: Silhouette scores computed from the embedding spaces generated by the proposed methods (HC, HCλ, HCE, HCEλ, RBL), as well as SupCon and OpenL3 reduced to 256 dimensions using PCA. The scores are evaluated across three dataset combinations: *Test*, *Test + U-ID* (test set combined with unseen individuals from known species), and *Test + U-species* (test set combined with unseen species). Silhouette scores are reported for each level of the hierarchy (ID, species, taxon) and averaged, with higher values indicating better cluster compactness and separation.

Dataset	Level	openl3 _pca256	SupCon	HC	HCλ	HCE	HCEλ	RBL
Test	ID	0.925	0.915	0.935	0.931	0.935	0.936	**0.949**
	Species	0.997	0.998	**0.999**	**0.999**	**0.999**	**0.999**	0.993
	Taxon	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**
U-ID (Novel ALL)	ID	0.936	0.892	0.882	0.909	0.728	0.907	**0.941**
	Species	0.991	0.994	**0.998**	**0.998**	0.995	0.997	0.983
	Taxon	0.998	0.996	**1.000**	0.999	0.998	0.998	**0.999**
U-ID (1-shot)	ID	0.243	0.197	0.203	0.242	0.121	0.230	**0.238**
	Species	0.951	0.972	**0.983**	0.979	0.966	0.977	0.874
	Taxon	0.987	0.988	**0.995**	0.990	0.984	0.990	0.976
U-ID (5-shot)	ID	**0.569**	0.481	0.463	0.538	0.311	0.526	0.559
	Species	0.967	0.984	**0.992**	0.988	0.986	0.988	0.917
	Taxon	0.992	0.992	**0.997**	0.996	0.996	0.996	0.995
U-species (Novel ALL)	ID	0.536	0.294	0.311	0.364	0.144	0.347	**0.567**
	Species	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Taxon	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
U-species (1-shot)	ID	0.968	0.944	0.952	**0.978**	0.580	0.975	0.821
	Species	0.968	0.944	0.952	0.978	**0.992**	0.975	0.820
	Taxon	0.987	0.988	0.990	**0.996**	0.993	0.988	0.992
U-species (5-shot)	ID	**1.000**	0.993	0.999	0.999	0.894	**1.000**	**1.000**
	Species	**1.000**	0.993	0.999	0.999	0.894	**1.000**	**1.000**
	Taxon	**1.000**	0.997	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 6.3: Accuracy results from kNN classification applied to the embedding spaces generated by the proposed approaches (HC, HCλ, HCE, HCEλ, and RBL), as well as SupCon and OpenL3 reduced to 256 dimensions using PCA. Evaluation is performed across the three datasets: *Test*, *U-ID* (unseen individuals from known species) and *U-species* (unseen individuals from novel species). For both U-ID and U-species datasets, three evaluation conditions are reported: Novel ALL (all available samples), 1-shot (single support example per class), and 5-shot (five support examples per class)

6.5 Discussion

A central objective of this chapter was to investigate if DBL methods could effectively be used to build hierarchically structured embedding spaces, and whether these spaces support generalisation to novel classes. In this section we discuss the main findings and interpret them in light of these goals.

Learning hierarchically structured embedding spaces: The silhouette score results in Table 6.2 show that HCL based methods consistently produce embedding spaces with more clearly defined clusters across all hierarchical levels. Among these, the **HCE** variant exhibits the strongest hierarchical structure which indicates that enforcing hierarchical consistency, by constraining the loss at each level on the loss value of the levels above, effectively promotes more structured and semantically meaningful embeddings. Another aspect captured

Dataset	Level	**OpenL3-32**	**HC$_{32}$**	**OpenL3-64**	**HC$_{64}$**	**OpenL3-128**	**HC$_{128}$**
Test	ID	**0.886**	0.390	**0.907**	0.860	**0.921**	0.908
	Species	**0.997**	0.960	0.997	**0.999**	0.997	**0.999**
	Taxon	**0.999**	0.983	0.999	**1.000**	0.999	**1.000**

Table 6.4: Ablation study on embedding dimensionality: kNN classification accuracy obtained from OpenL3 embeddings and HC models trained using embedding sizes of 32, 64, and 128 dimensions. Results are reported for the Test set at the ID, species, and taxon levels.

Dataset	Level	**OpenL3-32**	**HC$_{32}$**	**OpenL3-64**	**HC$_{64}$**	**OpenL3-128**	**HC$_{128}$**
Test	ID	**0.01**	-0.023	0.02	**0.237**	0.021	**0.236**
	Species	0.264	**0.549**	0.252	**0.458**	0.243	**0.396**
	Taxon	0.331	**0.516**	0.323	**0.408**	0.317	**0.346**
	AVERAGE	0.202	**0.348**	0.199	**0.368**	0.193	**0.326**

Table 6.5: Ablation study on embedding dimensionality: Silhouette scores obtained from OpenL3 embeddings and HC models trained using embedding sizes of 32, 64, and 128 dimensions. Results are reported for the Test set at the ID, species, and taxon levels, and on average across the levels.

on these results is the higher capacity of the **HCE** embeddings to represent novel classes, when compared with the other models. However overall, the silhouette scores computed when the U-id and U-species sets are added, indicate that the hierarchical structures are negatively impacted when the novel classes are introduced.

Reflecting on the operation of **SupCon**, which optimises exclusively for separation at the individual (ID) level, the fact that silhouette scores remain comparably high at the Species and Taxon levels can be interpreted as further evidence that our dataset naturally exhibits hierarchical structure. In other words, this suggests that mechanisms encoding individual identity in vocalisations may, to some extent, follow a phylogenetic process (as discussed in Section 2.1.2), leading to structured acoustic similarities across taxonomic levels.

The results obtained for **RBL** suggest that our implementation did not succeed in learning a meaningful hierarchical embedding space. When comparing the silhouette scores from this experiment with those obtained using the initial **OpenL3** embeddings, we observe minimal differences. This further supports the conclusion that **RBL** was not able to significantly reshape the initial embeddings to reflect the intended hierarchical structure.

Silhouette scores for the **OpenL3** embeddings reveal a lower degree of cluster definition at the individual (ID) level. However, scores for the Species and Taxon levels are more comparable to those of the other models. This pattern,

Dataset	Level	HCE	H-MTL
Test	ID	**0.196**	0.106
	Species	**0.394**	0.349
	Taxon	**0.363**	0.359
	Avg.	**0.318**	0.271
Test + U-id	ID	**0.130**	0.078
	Species	**0.338**	0.284
	Taxon	**0.250**	0.249
	Avg.	**0.239**	0.204
Test + U-species	ID	**0.159**	0.082
	Species	**0.374**	0.349
	Taxon	**0.311**	0.303
	Avg.	**0.281**	0.245

Table 6.6: Silhouette scores computed from the embedding spaces generated by HCE and H-MTL (hierarchical multitask learning approach) from the previous chapter (see Sec. 5.2.1). The scores are evaluated across three dataset combinations: *Test*, *Test + U-ID* (test set combined with unseen individuals from known species), and *Test + U-species* (test set combined with unseen species). Silhouette scores are reported for each level of the hierarchy (ID, species, taxon) and averaged, with higher values indicating better cluster compactness and separation.

supported by the UMAP visualisations in Figure. 6.6 suggests that the **OpenL3** model already exhibits a reasonable degree of species-level separation, possibly a result of its original training domain.

Finally, comparing learning paradigms and their ability to learn structured embedding spaces, Silhouette scores for **H-MTL** and **HCE** in Table.6.6, show the advantage of the DBL based method over the cross entropy based.

Structured embeddings as backbone of distance-based classification: We evaluated the suitability of the learned embedding spaces to serve as the backbone for distance-based classifiers by analysing KNN classification accuracy across the different models.

Results on the test set show that the **HCE**λ model offers a slight but consistent advantage at the ID level over all other models. Compared to **SupCon**, the higher accuracy values with **HCE**λ suggest that incorporating hierarchical structure in the embedding space can improve classification in the contrastive based approaches. These findings are also consistent with those in the previous chapter (see Section 5.5).

Among the HCL based models, λ versions show larger capacity to generalise to novel classes despite the general drop in accuracy.

Surprisingly, the **OpenL3** embeddings, despite lacking explicit hierarchical

Dataset	Level	HCE	H-MTL
Test	ID	0.930	**0.946**
	Species	**1.000**	**1.000**
	Taxon	**1.000**	**1.000**
U-id (Novel ALL)	ID	0.878	**0.919**
	Species	**1.000**	0.996
	Taxon	**1.000**	**1.000**
U-species (Novel ALL)	ID	0.461	**0.554**
	Species	**0.994**	0.977
	Taxon	**0.998**	0.992

Table 6.7: Comparison of accuracy results from kNN classification applied to the embedding spaces generated by HCE and H-MTL (hierarchical multitask learning approach) from the previous chapter (see Sec. 5.2.1). Evaluation is performed across the three datasets: *Test*, *U-ID* (unseen individuals from known species) and *U-species* (unseen individuals from novel species). For both U-ID and U-species datasets, we report results for the Novel ALL evaluation strategy, (all available samples as reference).

structure and being trained on a general audio tagging task, achieve comparable (and sometimes superior) KNN performance. And this is also true for the novel classes. This challenges the assumption that hierarchical structure is essential for classification in AIID and suggests that embeddings trained on large, diverse datasets can effectively separate classes without needing explicit embedding structures.

This raises the question: does **OpenL3** succeed simply because the embedding space is large enough to accommodate maximum separability between all classes even without a structured representation? In tables 6.4 and 6.5, we explore the impact of embedding dimensions on KNN accuracy and silhouette scores in OpenL3 and HC models. We can observe that classification performance drops for both **OpenL3** and **HCL** models when dimensionality is reduced to 32, but OpenL3 remains relatively robust. This indicates that OpenL3 can maintain class separability even in low-dimensional settings, while contrastive models struggle to preserve fine-grained distinctions, particularly when also pushing for a hierarchical alignment.

Interestingly, both evaluation metrics (silhouette scores and KNN accuracy) show a complementary pattern, i.e. the models that show a more hierarchically structured embedding space are not necessarily the ones that show the best separability between classes and thus accuracy. The accuracy results for **RBL** reinforce this idea as well. Taken together, these findings suggest that while hierarchical supervision enhances embedding structure, its advantages for classification are not consistently demonstrated. The effectiveness of such structure may depend on other factors, including the dimensionality of the embedding

space and the nature of the pretraining domain.

Notably, pretrained models like OpenL3 can achieve competitive classification performance despite lacking explicitly structured or semantically aligned embeddings. This aligns with the earlier observation that RBL did not substantially alter the original embedding space. The relatively strong classification accuracy obtained with RBL may reflect a balance between the latent structure already present in the pretrained embeddings and the limited contribution of the hierarchical supervision applied during training.

An important factor to consider when comparing these models is the distance metric used during both training and evaluation. While OpenL3 and HC embeddings are trained using a cosine similarity objective, RBL embeddings are optimized based on pairwise Euclidean distances. Moreover, the evaluation metrics in our experiments are also computed using Euclidean distance. We hypothesise that this mismatch between the training and evaluation metrics across OpenL3, HC, and RBL may have contributed to some of the unexpected results observed. In particular, the limited alteration of the embedding space by RBL relative to the OpenL3 baseline may stem from this inconsistency. Since the embedding geometry is inherently shaped by the distance metric used during training, evaluating under a different metric can distort the neighbourhood structure of the space, leading to discrepancies in both classification accuracy and silhouette scores.

This insight challenges our earlier interpretation that hierarchical structuring of the embedding space is not beneficial, especially when the backbone representation is already strong, as in OpenL3. Instead, these findings suggest the need for an explicit analysis of the embedding geometry and for ensuring metric consistency across training and evaluation to enable fair and meaningful comparisons between models.

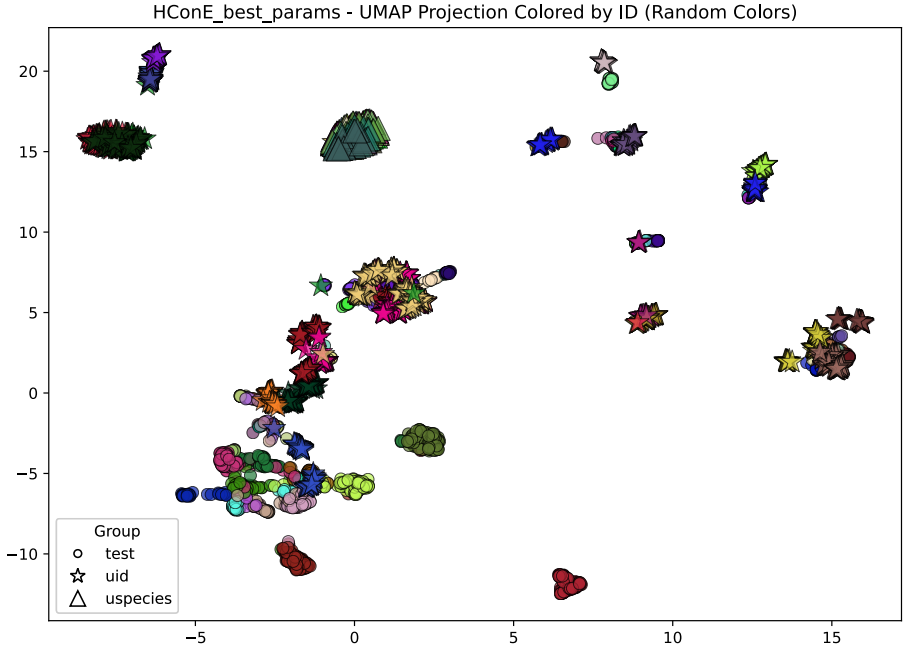
6.6 Conclusion

In this chapter, we explored whether DBL methods can be used to effectively construct hierarchically structured embedding spaces and whether these representations support generalisation to novel classes.

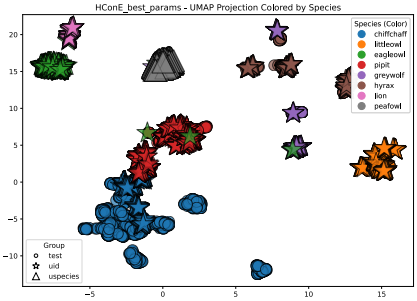
The proposed approaches, hierarchical contrastive learning and rank based loss, were evaluated against the flat contrastive approach (SupCon) and the OpenL3 model. Results show that hierarchical supervision through a contrastive process improves the hierarchical organisation of the embedding space, as reflected by higher silhouette scores across taxonomic levels. However, improvements in structure do not always translate into better classification accuracy. Among all models, the best-performing systems appear to establish a balance

between the inherent discriminative power of the initial embeddings and the hierarchical supervision applied during training. This is evident in the performance of RBL, which did not drastically reshape the original embedding space but still achieved competitive results in classification.

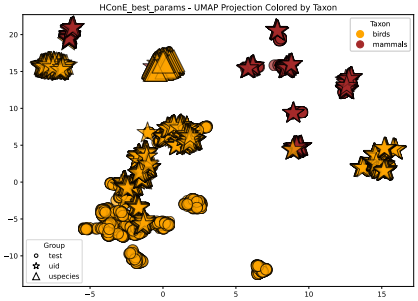
Finally, it it important to reflect on the limitations of evaluating embedding spaces solely through the proposed silhouette scores, kNN accuracy, and UMAP visualisations. While these metrics can indicate aspects related to the structure of the embedding spaces, each presents specific challenges that limit their interpretability. A future direction of this study should be on establishing strategies for evaluation of embedding spaces.



(a) HCEλ - ID level

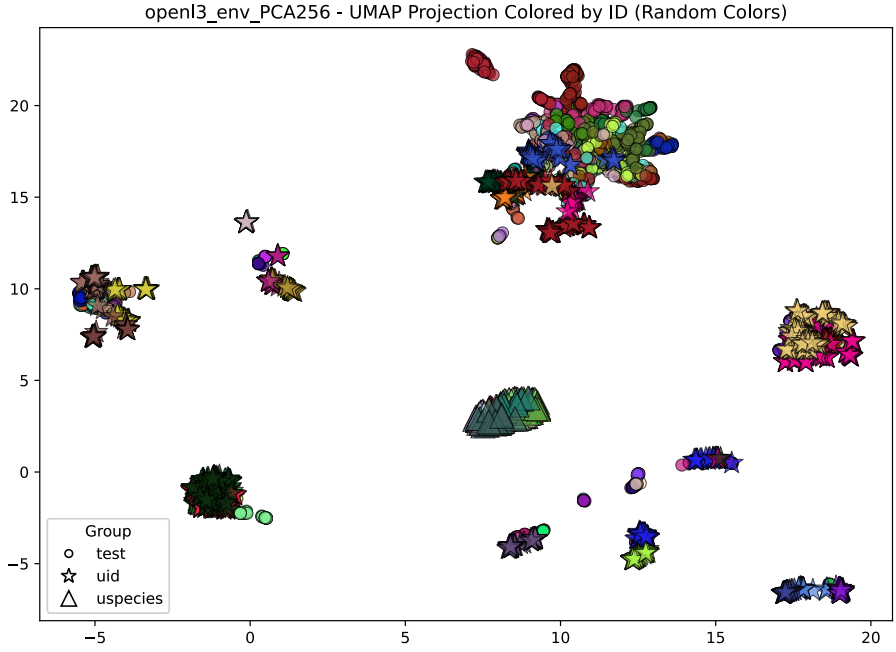


(b) HCEλ - Species level

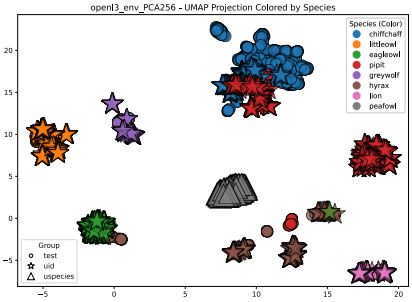


(c) HCEλ - Taxon level

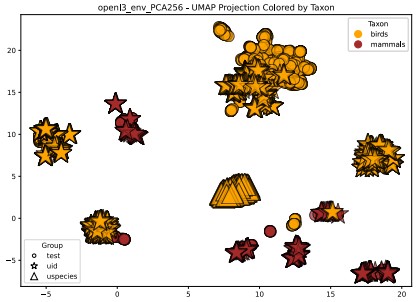
Figure 6.5: UMAP visualisation of the embedding spaces produced by the **HCE**λ model across all evaluation sets. The top panel shows the individual (ID) level; the bottom panels show species and taxon levels. Colours indicate class membership at each respective level, while marker shapes denote the evaluation subset (test, unseen individuals, or unseen species). These plots highlight differences in embedding structure and class separability.



(a) openl3_PCA256 ID level



(b) openl3_PCA256 Species level



(c) openl3_PCA256 Taxon level

Figure 6.6: UMAP visualisation of the embedding spaces produced by the **openl3_PCA256** model across all evaluation sets. The top panel shows the individual (ID) level; the bottom panels show species and taxon levels. Colours indicate class membership at each respective level, while marker shapes denote the evaluation subset (test, unseen individuals, or unseen species). These plots highlight differences in embedding structure and class separability.

Chapter 7

Conclusion and Final remarks

This thesis set out to develop systems for Acoustic Identification of Individual Animals (AIID) that are capable of operating under real-world conditions. In framing the problem for such settings, we identified two key research questions that have guided this work:

Multispecies AIID: Can a single AIID model generalise to multiple species across taxonomic groups?

Generalisation to novel classes: Can such a system be built that is able to generalise to novel individuals and species not seen during training?

In this final chapter, we review the main ideas explored and reflect on the broader significance of our findings. Section 7.1 provides an overview of our approach and the major outcomes of the work. Section 7.2 discusses the implications of our methodological choices and findings. We focus on key themes such as data representation, hierarchical framing of AIID and generalisation. Section 7.3 outlines future directions and open challenges, including the potential for open world classification. Finally, Section 7.4 concludes with broader reflections on the contributions of this work.

7.1 Overview

To address the research questions outlined above, we began by proposing a shift from the traditional approach of training separate, species-specific AIID models to a unified method that can operate across diverse taxa. To support this development we curated a multispecies dataset containing vocalisations from

both mammals and birds and defined 3 evaluation sets. These were designed to assess model performance both on known classes (seen during training) and on novel classes with varying levels of difficulty, as detailed in Chapter 4.

Inspired by the parallels between the multispecies AIID problem and the few-shot bioacoustic event detection task proposed in the DCASE challenge (see Section 3.4.1), we leveraged the representational power of pretrained models to serve as backbone feature extractors for our systems. To inform the selection of a suitable pretrained model, we conducted an exploratory study on data representation, presented in Section 4.3. Based on this analysis, the OpenL3 model was selected for its favourable performance across species and classification levels.

A hypothesis we follow in this work is that knowledge sharing across related tasks is beneficial towards our multispecies AIID goals. This is explored in Chapter 5. Transference of knowledge is present here in two ways: across species and across hierarchical classification tasks (Taxon, species and ID classification). From this analysis comes an important outcome of this work: learning classifiers by including hierarchically related tasks is consistently advantageous compared with flat approaches, i.e that only address the ID task.

To directly address generalisation to novel classes, and again influenced by our study in chapter 3, we proceeded to explore distance based learning approaches that work towards learning embedding spaces particularly suitable for the multispecies AIID problem in chapter 6. By including the hierarchical structure as a training goal we were able to develop methods that create a single embedding space that can (in the same space) represent a multitude of classes, at different levels of granularity and preserve hierarchic consistency. Whether such structures show an advantage in the generalisation to novel classes, the response is less clear and results suggest that other aspects such as dimension and training domain of the initial representations impact these approaches considerably.

7.2 Discussion

This discussion is structured around a series of key methodological choices that shaped the development of this thesis. We reflect on the validity of these decisions and their impact on the proposed approaches to multispecies AIID.

7.2.1 Framing multispecies AIID as a set of small related tasks within a unified model

Common approaches in bioacoustics either consider tasks as one big homogeneous problem, that can be solved by supervised learning approaches supported by large enough data, or they break the problem into smaller disjoint tasks that can be solved individually by highly specialised models. Motivated by the work described in chapter 3, we instead frame multispecies AIID as a collection of smaller tasks that are related to each other and can be solved through a unified approach.

We find this framing to be beneficial. As described in Chapter 4, the multispecies AIID dataset is composed of multiple species-specific subsets that differ significantly in both acoustic characteristics and data volume. This confirms that the overall dataset is not homogeneous and cannot be treated as a single, uniform classification problem. At the same time, the species are not unrelated, since there are meaningful biological and acoustic relationships across them, which justifies treating them as a set of disjoint but related tasks. This view is further supported by our experiments in chapter 5. The flat approach (MS-AIID) which treats the problem as a single undifferentiated task results in poor performance, while Multi-task approaches (MTL) that leverage the taxonomic relationships across tasks perform better.

Additionally, this framing aligns with the second goal of the thesis: enabling generalisation to novel classes. By adopting a unified approach that leverages relationships across tasks, we create the conditions necessary to reason beyond the specific individuals and species present in the training data. This perspective motivates the distance-based approaches proposed in Chapter 6, which aim to learn a single embedding space capable of representing a wide range of distinct classes. While full generalisation to unseen classes remains a challenge, this formulation of the problem is what makes such generalisation a possible objective.

7.2.2 Using pretrained embeddings as a representation backbone

In low-data scenarios, it is common practice to leverage the representational power of large models pretrained on extensive datasets to encode the input data. One particularly attractive feature of using pretrained embeddings is the consistent way in which they represent a wide variety of examples. Chapter 3 motivates and exemplifies this practice, highlighting the importance of standardising input representations across acoustic signals with widely varying

characteristics and durations.

An exploratory study on data representation, presented in Chapter 4, further demonstrated that certain pretrained models can capture the structure of our dataset without losing the fine-grained information necessary for distinguishing between individuals. Based on this analysis, the OpenL3 model, trained on environment sounds, was selected as the primary embedding extractor used throughout this thesis. Its suitability was further validated in Chapter 6, where it showed strong performance at the multispecies AIID task when coupled with simple KNN classifiers.

This practice comes with some disadvantages as well. First, it decouples the feature extraction process from the specific classification task, making it difficult to associate model performance with specific acoustic characteristics of the input signals. As a result, interpretability suffers. In our case, adopting pretrained embeddings limits our ability to connect findings with domain knowledge on acoustic signatures. Given the biological relevance of such information, particularly in studies of individual vocal identity, we consider this to be limitation of the work.

A second concern is the potential introduction of unintended biases stemming from the domain in which the pretrained model was originally trained. In Chapter. 4 we show the variation in results obtained when changing the training domain of the pretrained model, (openl3_music vs Openl3_Env). While the embeddings serve as a strong initialisation for downstream learning, it is difficult to characterise precisely how this pretraining influences performance. The representational power of pretrained models like OpenL3 can vary considerably across application domains, yet this is not possible to measure without prior experimentations.

7.2.3 Animal taxonomy as a hierarchical structure for multispecies AIID

The main way in which we relate species-specific AIID problems to one another is by framing them through a hierarchical structure that mirrors animal taxonomy, including taxon, species, and individual identity. Instead of treating classification as a flat problem, we construct three classification tasks that reflect biologically meaningful relationships across levels of organisation. Support for this framing comes from the phylogenetic considerations discussed in Section 2.1.2, which suggest that animal vocalisations are shaped by evolutionary pressures and often carry a phylogenetic signal. As a result, genetically related species are more likely to share acoustic traits than distantly related ones. However, this assumption does not hold universally. As discussed in our review

of acoustic signatures across taxa (Section 2.2.1), there are notable examples of closely related species with very different ways of encoding identity in their vocalisations, which are often due to ecological, or environmental adaptations. This raises important questions about the reliability of using taxonomic distance as a predictor of vocal similarity.

Despite these, in our work, incorporating taxonomic structure into the models has produced clear advantages. For instance in chapter 5, a multi-task learning approach that jointly optimises classification across taxon, species, and individual levels outperforms flat approaches that focus on identity alone. These results support the conclusion that hierarchical structure can effectively model the species present in our dataset. Nevertheless, looking towards scaling these methods to a broader range of taxa, we recognise the need for better tools to identify and accommodate exceptions to the assumed taxonomic structure.

We believe that the choice to use animal taxonomy as the hierarchical structure of the problem is well motivated, as it provides a biologically grounded framework that captures evolutionary relationships. However, this approach can be seen as fitting the problem into a predefined mould, which may not fully reflect the acoustic similarities that are most relevant for vocal identification. In the future, it would be valuable to explore alternative structures driven primarily from the similarity of acoustic features in vocalisations themselves. For example, Odom et al. [2021] discusses approaches for comparing vocalisations and deriving phylogenetic trees based directly on acoustic characteristics. Extending this idea further, one possible direction would be to learn such acoustic hierarchies directly from the data in an unsupervised manner, uncovering structures based on acoustic similarity. These learned hierarchies could potentially serve as a more suitable backbone for constructing AIID classifiers and aid on building domain knowledge regarding acoustic signatures.

7.2.4 Generalisation to unseen classes through distance-based learning

The decision to adopt a DBL paradigm for multispecies AIID was motivated by the demonstrated ability of DBL methods to construct continuous, semantically meaningful embedding spaces. Unlike classification models constrained to learn representations of a fixed label set, DBL approaches aim to learn a representational space in which novel classes can be naturally embedded, based on their semantic similarity to previously seen examples. This property is essential for the ability to generalise to unseen individuals or species.

To assess the quality of the learned embedding spaces, we evaluated how well they express the underlying hierarchical structure of the data. Specifically, we

computed silhouette scores at multiple taxonomic levels to capture how tightly grouped the embeddings are by individual, species, and taxon. As exemplified in Table 6.6, HCE (a hierarchical contrastive DBL method) consistently produced more structured embeddings than cross-entropy-based methods such as H-MTL, even when novel classes are projected into the space. These findings validate the original hypothesis that DBL methods are advantageous in learning representations that can generalise to novel classes.

However, an important finding is that a structural advantage does not always translate into superior classification accuracy. As seen in Table 6.7, the KNN results show that HCE fails to outperform H-MTL even on novel classes. This disconnect between embedding quality and predictive performance is consistently present throughout the experiments in Chapter 6 and can be partially attributed to the influence of the pretrained initialisation.

Evaluation of the frozen OpenL3 embeddings confirms that it already provides a highly separable space for individual identity. When we apply H-MTL and RBL approaches onto these initialisation, the methods only shape this space lightly. By contrast, the hierarchical-contrastive based methods such as HCE appear to reshape the embedding geometry far more aggressively in pursuit of the targeted hierarchical structure. In doing so, they may undo much of the fine-scale structure inherited from the OpenL3 initialisation, which could explain the poorer KNN accuracy observed. It is possible that the results we report are simply the result of an incomplete training process. Perhaps with a longer optimisation process, these space would converge to a configuration that retains both taxonomic coherence and local separability. This perspective also aids in interpreting the results obtained with RBL. Instead of an aggressive reshaping of the initial space, because its target distances are derived directly from the pairwise relations that already exist in the OpenL3 space, RBL nudges the embeddings toward a more hierarchical arrangement. As a result, it preserves much of OpenL3's original class-separating capability while still improving global taxonomic consistency.

7.3 Limitations and future work

In working towards multispecies AIID systems that can be deployed in the wild, this work necessarily made a number of simplifications. While these decisions were important to isolate contributions from the proposed approaches, they also leave several challenges unaddressed, particularly those relating to the complexity of the data.

As already identified in Chapter 4, there are two key simplifications in the dataset that may affect generalisation. First, we do not address the presence

of confounding factors, such as environmental acoustic characteristics or artefacts that consistently co-occur with specific individuals, and can lead models to associate incidental cues with identity rather than learning vocal features. Second, we assume that each animal has a single and stable acoustic signature, ignoring the possibility that these signatures may change over time or vary across call types (see Section 2.2.1). These assumptions may compromise model performance in different scenarios than the ones present in training data.

To address these limitations, future work should aim to increase variation in the training data by including recordings across different environments, seasons, and recording devices, as well as by explicitly incorporating multiple call types per individual. In line with the the work by Stowell et al. [2016], stratified data augmentation could be employed to mix background sounds across individuals, reducing the reliance on spurious correlations. Extending the dataset to include recordings of the same individuals under varied conditions along with background audio, would provide the necessary support to mitigate this issues.

Contrary to these simplifications, the unbalanced nature of the dataset is a complexity that we maintained. This decision was motivated by the desire to approximate real world conditions, however, in hindsight, this introduced additional challenges that limited the success of certain approaches explored. The training of the MS-MTL model was largely unsuccessful, primarily due to the difficulty of balancing the contribution of each species' data during network optimization. Moreover, dataset imbalance complicates the evaluation of embedding spaces since it affects KNN classification outcomes and makes the interpretation of the resulting geometry more uncertain.

During dataset construction, we were constrained by the availability of recordings, which limited our control over species selection. Although we aimed to maintain a roughly even representation between bird and mammal species, we did not regulate factors such as phylogenetic relationships, sex, or age of the individuals included. These aspects, among others, likely influence the geometry of the hierarchical embedding space by introducing additional sources of similarity and dissimilarity among vocalisations.

As expected, the idealised hierarchical embedding space illustrated in Fig. 6.1 is overly simplified since it assumes phylogeny as the dominant structuring factor while neglecting other biological or contextual influences. Without control over these variables, the interpretation of the learned embedding spaces becomes more difficult.

Future work should therefore prioritise careful curation of the species and individuals included, ensuring diversity in relevant biological and ecological characteristics. Constructing both balanced and intentionally unbalanced dataset versions would further support a more systematic understanding of how data

composition affects the behaviour of the models and the geometry of the learned embedding spaces. At the time of writing, such a dataset still does not exist. Research on AIID remains largely focused on single-species settings, leaving the potential benefits of jointly training across multiple species under explored. Although interest in applying machine learning to AIID has grown and several new datasets have been released, they continue to target individual species rather than providing a unified, multispecies resource. Datasets such as Martin et al. [2022b] or even the datasets in Hagiwara et al. [2023a] which provide individual labels, while valuable, still fit within this single-species paradigm. However these are important sources that could be incorporated into a broader, standardised hierarchical AIID dataset in the future.

During the development of this work, we identified two promising directions that could not be fully pursued. The first is to frame multispecies AIID as a few-shot task. Given the overlap with the few-shot bioacoustic event detection task explored in Chapter 3, we could explore techniques such as prototypical networks. For instance, the work by Liang et al. [2022], that integrates hierarchical structures into the few-shot learning paradigm, provides an interesting approach. More recently, the rise of multimodal foundation models has presented powerful ways for tackling few-shot acoustic identification. Specifically, models adapted from Large Language Models (LLMs), such as Contrastive Language-Audio Pretraining (CLAP)[Elizalde et al., 2023], establish a shared embedding space between audio and text representations. This enables the application of powerful LLM capabilities, like in-context learning, to acoustic tasks. An extreme case of few-shot learning, zero-shot learning, is now being explored by these systems (e.g., NatureLLM-Audio [Robinson et al., 2024]). In zero-shot identification, the system can classify a novel individual or species purely based on a textual description or label, without having observed any corresponding acoustic samples during the query phase, significantly reducing the need for labelled training data.

The second direction involves directly incorporating domain knowledge about acoustic signatures into the design of the model. This approach is based in the assumption that vocal individuality arises from a limited set of mechanisms, such as specific spectral features or temporal patterns. and that these may also be structured in a similar way as animal taxonomy. Extending on our hierarchical classification approach, we could design a system that guides the decisions across a sequence of hierarchically related tasks such as taxon, species, AIID strategy, and individual identity. By progressing through this hierarchy, the model could progressively narrow its focus, attending to the most relevant parts of the signal at each level. The attraction of such solution is that it is inherently interpretable and has the potential to reveal high-level insights into the nature

of vocal individuality across taxa.

Finally, one of the main goals of this thesis — generalisation to novel classes — was motivated by the broader challenge of performing open world AIID classification. In real-world scenarios, the ability to identify and incorporate previously unseen individuals or species is essential. While our results demonstrate some ability to meaningfully represent novel classes, we did not discuss how that can fit into a complete pipeline for open world classification. Such a system must go beyond simply embedding novel samples: it must also include mechanisms for detecting out-of-distribution inputs [Hogeweg et al., 2024], reasoning over hierarchical taxonomies, and incorporating new classes into the label structure over time. The overarching goal in this context is for each vocalisation to be classified to the maximum extent of what is known to the model. That is, even when a vocalisation from a novel individual is encountered, the system should be able to recognise it as a new individual of a known species. If the vocalisation comes from an entirely unknown species, the model should at least be able to indicate its broader taxonomic group, such as identifying it as a bird or a mammal. This incremental mechanism aligns with the hierarchical structure of the target embedding spaces and provides a concrete way forward to realise open world AIID.

7.4 Final Remarks

In this work, we explored machine learning methods for the problem of acoustic identification of individual animals, guided by the two central goals stated at the beginning of this chapter: enabling multispecies identification and supporting generalisation to novel classes.

As expected, multispecies AIID proved to be a challenging task, requiring approaches that go beyond traditional supervised classification. In our work using multitask learning we demonstrated that it is possible to perform multispecies AIID on a limited but taxonomically diverse dataset. These results provide promising evidence that a unified model can effectively operate across multiple species.

The second step focused on the ability to effectively classify the novel classes. Overall the accuracy results on the novel classes remain below 50% across all models. Moreover, it remains difficult to disentangle the effects of dataset-specific factors from those of the pretrained embeddings used throughout as data representation. Notably, we observed an inverse relationship between the quality of hierarchical structure in the embedding space and the classification accuracy achieved on novel classes. This results warrants further investigation to determine whether it reflects a fundamental principle or simply arises from

limitations in the training process or implementation details.

To finalise, it is remarkable to witness how bioacoustics has recently gained significant momentum, driven by the rapid transfer of knowledge from computer science into ecological and behavioural research. This cross-disciplinary exchange is already transforming the field, enabling analyses and insights that were, until recently, purely aspirational.

It is an exciting time to contribute to AIID research, as we move from speculative ideas to tangible goals such as approaching the possibility of decoding animal communication or identifying individuals in the wild and count individuals. We are indeed on the verge of what once seemed beyond reach just a few years ago.

At the same time, we must remain aware that our work concerns other living beings whose communication systems and social worlds we still understand only partially. As the field advances, we need to place greater emphasis on integrating behavioural and ecological knowledge into computational approaches. This requires close collaboration with the researchers who study animal communication and, as far as possible, with the animals themselves, ensuring that our methods do not cause unintended harm. In doing so, we can ensure that advances in AIID are not only on the side of machine learning but also foster a more supported and biologically informed understanding of animal communication.

Bibliography

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Kuntoro Adi, Michael T. Johnson, and Tomasz S. Osiejuk. Acoustic censusing using automatic vocalization classification and identity recognition. *The Journal of the Acoustical Society of America*, 127(2):874–883, 2010. ISSN 0001-4966. doi: 10.1121/1.3273887.

Jozsef Arato and W Tecumseh Fitch. Phylogenetic signal in the vocalizations of vocal learning and vocal non-learning birds. *Philosophical transactions of the royal society B*, 376(1836):20200241, 2021.

David W. Armitage and Holly K. Ober. A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics*, 5(6):465–473, 2010. ISSN 15749541. doi: 10.1016/j.ecoinf.2010.08.001. URL `http://dx.doi.org/10.1016/j.ecoinf.2010.08.001`.

MAO Axiu, Claire Giraudet, LIU Kai, Ines De Almeida Nolasco, Zhiqin Xie, Zhixun Xie, Yue Gao, James Theobald, Devaki Bhatta, Rebecca Stewart, et al. Automated identification of chicken distress vocalisations using deep learning models. *bioRxiv*, 2021.

Carol L Bedoya and Laura E Molles. Acoustic censusing and individual identification of birds in the wild. *bioRxiv*, pages 2021–10, 2021.

Michael D Beecher. Signalling systems for individual recognition: an information theory approach. *Animal Behaviour*, 38(2):248–261, 1989.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Abhijit Bendale and Terrance Boult. Towards Open World Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision*

and Pattern Recognition, volume 07-12-June, pages 1893–1902, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298799.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 1563–1572, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.173.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13979–13988, 2021.

Holly A Broadhurst, Naiara Guimarães Sales, Robert Raynor, Claire Howe, Erinma Ochu, Xavier Lambin, Christopher S Sutherland, and Allan D McDevitt. From water to land: A review on the applications of environmental dna and invertebrate-derived dna for monitoring terrestrial and semi-aquatic mammals. *Mammal Review*, page e70006, 2025.

Nora V Carlson, E McKenna Kelly, and Iain Couzin. Individual vocal recognition across taxa: a review of the literature and a look into the future. *Philosophical Transactions of the Royal Society B*, 375(1802):20190479, 2020.

Alecia J Carter, Harry H Marshall, Robert Heinsohn, and Guy Cowlishaw. Personality predicts the propensity for social learning in a wild primate. *PeerJ*, 2:e283, 2014.

Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5953–5963, 2024.

Isabelle Charrier, Benjamin J Pitcher, and Robert G Harcourt. Vocal recognition of mothers by australian sea lion pups: individual signature and environmental constraints. *Animal Behaviour*, 78(5):1127–1134, 2009.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020b.

Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, Trevor Darrell, et al. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2 (3):5, 2020c.

Jinkui Cheng, Yuehua Sun, and Liqiang Ji. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. *Pattern Recognition*, 43(11):3846–3852, 2010. ISSN 00313203. doi: 10.1016/j.patcog.2010.04.026. URL `http://dx.doi.org/10.1016/j.patcog.2010.04.026`.

Christopher James Clark. *Locomotion-Induced Sounds and Sonations: Mechanisms, Communication Function, and Relationship with Behavior*, pages 83–117. Springer International Publishing, Cham, 2016. ISBN 978-3-319-27721-9. doi: 10.1007/978-3-319-27721-9_4. URL `https://doi.org/10.1007/978-3-319-27721-9_4`.

Aurora Linh Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE, 2019.

Jason Cramer, Vincent Lostanlen, Andrew Farnsworth, Justin Salamon, and Juan Pablo Bello. Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 901–905, 2020. doi: 10.1109/ICASSP40776.2020.9052908.

Leonidas-Romanos Davranoglou, Graham K Taylor, and Beth Mortimer. Sexual selection and predation drive the repeated evolution of stridulation in heteroptera and other arthropods. *Biological Reviews*, 98(3):942–981, 2023.

DCASE. DCASE challenge 2022 Few-shot bioacoustic event detection task - results page, 2022. URL `https://dcase.community/challenge2022/task-few-shot-bioacoustic-event-detection-results`. Accessed: 2022-09-27.

DCASE. DCASE challenge 2023 Few-shot bioacoustic event detection task - results page, 2023. URL `https://dcase.community/challenge2023/task-few-shot-bioacoustic-event-detection-results`. Accessed: 2023-07-20.

María del Mar Delgado, Eleonora Caferri, Maria Mendez, Jose A Godoy, Letizia Campioni, and Vincenzo Penteriani. Population characteristics may reduce the levels of individual call identity. *PloS one*, 8(10):e77557, 2013.

Anthony I Dell, John A Bender, Kristin Branson, Iain D Couzin, Gonzalo G de Polavieja, Lucas PJJ Noldus, Alfonso Pérez-Escudero, Pietro Perona, Andrew D Straw, Martin Wikelski, et al. Automated image-based tracking and its application in ecology. *Trends in ecology & evolution*, 29(7):417–428, 2014.

Vlad Demartsev, Michal Haddas-Sasson, Amiyaal Ilany, Lee Koren, and Eli Geffen. Male rock hyraxes that maintain an isochronous song rhythm achieve higher reproductive success. *Journal of Animal Ecology*, 92(8):1520–1531, 2023.

Julie E. Elie and Frédéric E. Theunissen. Zebra finches identify individuals using vocal signatures unique to each call type. *Nature Communications*, 9 (1), 2018. ISSN 20411723. doi: 10.1038/s41467-018-06394-9.

Benjamin Elizalde, Abelino Jimenez, and Bhiksha Raj. Sound event classification using ontology-based neural networks. In *NIPS 2018 Workshop*, 2018.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Marco Gamba, Livio Favaro, Alessandro Araldi, Valentina Matteucci, Cristina Giacoma, and Olivier Friard. Modeling individual vocal differences in group-living lemurs using vocal tract morphology. *Current zoology*, 63(4):467–475, 2017.

Hugo Flores Garcia, Aldo Aguilar, Ethan Manilow, and Bryan Pardo. Leveraging hierarchical structures for few-shot musical instrument recognition. *arXiv preprint arXiv:2107.07029*, 2021.

Maxime Garcia and Livio Favaro. Animal vocal communication: function, structures, and production mechanisms, 2017.

Femke Gelderblom, Benjamin Cretois, Pal Johnsen, Filippo Remonato, and Tor Arne Reinen. Few-shot bioacoustic event detection using beats. Technical report, DCASE2023 Challenge, June 2023.

Katherine E Gentry, Rebecca N Lewis, Hunter Glanz, Pedro I Simões, Árpád S Nyári, and Michael S Reichert. Bioacoustics in cognitive research: Applications, considerations, and recommendations. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(5):e1538, 2020.

H Carl Gerhardt. The evolution of vocalization in frogs and toads. *Annual review of ecology and systematics*, pages 293–324, 1994.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

Franz Goller. The syrinx. *Current Biology*, 32(20):R1095–R1100, 2022.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Thibault Grava, Nicolas Mathevon, Emelyne Place, and Patrick Balluet. Individual acoustic monitoring of the european eagle owl bubo bubo. *Ibis*, 150 (2):279–287, 2008.

Jon Grinnell and KAREN McCOMB. Roaring and social communication in african lions: the limitations imposed by listeners. *Animal Behaviour*, 62(1): 93–98, 2001.

María José Guerrero Muriel, Carol Bedoya Acevedo, José David López Hincapié, Claudia Victoria Isaza Narváez, and Juan Manuel Daza Rojas. Acoustic animal identification using unsupervised learning. 2023.

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. Beans: The benchmark of animal sounds. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023a. doi: 10.1109/ICASSP49357.2023.10096686.

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. Beans: The benchmark of animal sounds. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023b.

Jenny Hamer, Eleni Triantafillou, Bart Van Merriënboer, Stefan Kahl, Holger Klinck, Tom Denton, and Vincent Dumoulin. Birb: A generalization benchmark for information retrieval in bioacoustics. *arXiv preprint arXiv:2312.07439*, 2023.

Loïc A Hardouin, Pierre Tabel, and Vincent Bretagnolle. Neighbour–stranger discrimination in the little owl, athene noctua. *Animal Behaviour*, 72(1): 105–112, 2006.

Luis Hernandez-Miranda and Carmen Birchmeier. Mechanisms and neuronal control of vocalization in vertebrates. *Opera Medica et Physiologica*, 4:50–62, 09 2018. doi: 10.20388/omp2018.001.0059.

Michael Hertkorn. Few-shot bioacoustic event detection : Don't waste information. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Hertkorn_28_5.pdf`.

Laurens E Hogeweg, Rajesh Gangireddy, Django Brunink, Vincent J Kalkman, Ludo Cornelissen, and Jacob W Kamminga. Cood: Combined out-of-distribution detection using multiple measures for anomaly & novel class detection in large-scale hierarchical classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2024.

Lifi Huang, Rohan Clarke, Daniella Teixeira, André Chiaradia, and Bernd Meyer. Acoustic recognition of individual animals in the presence of unknown individuals. *bioRxiv*, pages 2024–12, 2024.

Qisheng Huang, Yanxiong Li, Wenchang Cao, and Hao Chen. Few-shot bio-acoustic event detection based on transductive learning and adapted central difference convolution. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Huang_59_5.pdf`.

Stephen J Insley. Long-term vocal recognition in the northern fur seal. *Nature*, 406(6794):404–405, 2000.

Arindam Jati, Naveen Kumar, Ruxin Chen, and Panayiotis Georgiou. Hierarchy-aware Loss Function on a Tree Structured Label Space for Audio Event Detection. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019-May:6–10, 2019. ISSN 15206149. doi: 10.1109/ICASSP.2019.8682341.

Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

Ammie K. Kalan, Roger Mundry, Oliver J.J. Wagner, Stefanie Heinicke, Christophe Boesch, and Hjalmar S. Kühl. Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. *Ecological Indicators*, 54:217–226, 2015. ISSN 1470160X. doi: 10.1016/j.ecolind.2015.02.023. URL `http://dx.doi.org/10.1016/j.ecolind.2015.02.023`.

Taein Kang. Few-shot bioacoustic event detection using good embedding model. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Kang_10_5.pdf`.

Vincent S. Kather, Burooj Ghani, and Dan Stowell. Clustering and novel class recognition: evaluating bioacoustic deep learning feature extractors, 2025. URL `https://arxiv.org/abs/2504.06710`.

Arik Kershenbaum, Çağlar Akçay, Lakshmi Babu-Saheer, Alex Barnhill, Paul Best, Jules Cauzinille, Dena Clink, Angela Dassow, Emmanuel Dufourq, Jonathan Growcott, et al. Automatic detection for bioacoustic research: a practical guide from and for biologists and computer scientists. *Biological Reviews*, 100(2):620–646, 2025.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673, 2020.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. IEEE, 2018.

Elly Knight, Tessa Rhinehart, Devin R de Zwaan, Matthew J Weldy, Mark Cartwright, Scott H Hawley, Jeffery L Larkin, Damon Lesmeister, Erin Bayne, and Justin Kitzes. Individual identification in acoustic recordings. *Trends in Ecology & Evolution*, 39(10):947–960, 2024.

Lee Koren and Eli Geffen. Individual identity is communicated through multiple pathways in male rock hyrax (procavia capensis) songs. *Behavioral Ecology and Sociobiology*, 65:675–684, 2011.

Friedrich Ladich and Hans Winkler. Acoustic communication in terrestrial and aquatic vertebrates. *Journal of Experimental Biology*, 220(13):2306–2317, 2017.

Paola Laiolo. The emerging significance of bioacoustics in animal species conservation. *Biological conservation*, 143(7):1635–1645, 2010.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934, 2020.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

Yuna Lee, HaeChun Chung, and JaeHoon Jung. Few-shot bioacoustic detection boosting with fine tuning strategy using negative based prototypical learning. Technical report, DCASE2023 Challenge, June 2023.

Kenna DS Lehmann, Frants H Jensen, Andrew S Gersick, Ariana Strandburg-Peshkin, and Kay E Holekamp. Long-distance vocalizations of spotted hyenas contain individual, but not group, signatures. *Proceedings of the Royal Society B*, 289(1979):20220548, 2022.

Thierry Lengagne, Jacques Lauga, and Thierry Aubin. Intra-syllabic acoustic signatures used by the king penguin in parent-chick recognition: An experimental approach. *Journal of Experimental Biology*, 204(4):663–672, 2001. ISSN 00220949.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1446–1455, 2019.

Ren Li, Jinhua Liang, and Huy Phan. Few-shot bioacoustic event detection using prototypical networks with resnet classifier technical report. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Li_90_5.pdf`.

Jinhua Liang, QH Phan, Emmanouil Benetos, et al. Leveraging label hierarchies for few-shot everyday sound recognition. 2022.

Jinhua Liang, Ines Nolasco, Burooj Ghani, Huy Phan, Emmanouil Benetos, and Dan Stowell. Mind the domain gap: a systematic analysis on bioacoustic sound event detection. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1257–1261. IEEE, 2024.

Pavel Linhart, Hans Slabbekoorn, and Roman Fuchs. The communicative significance of song frequency and song length in territorial chiffchaffs. *Behavioral Ecology*, 23(6):1338–1347, 2012.

Pavel Linhart, Tomasz S. Osiejuk, Michał Budka, Martin Šálek, Marek Špinka, Richard Policht, Michaela Syrová, and Daniel T. Blumstein. Measuring individual identity information in animal signals: Overview and performance of available identity metrics. *Methods in Ecology and Evolution*, 10(9):1558–1570, 2019. ISSN 2041210X. doi: 10.1111/2041-210X.13238.

Haohe Liu, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley. Surrey system for dcase 2022 task 5 : Few-shot bioacoustic event detection with segment-level metric learning. Technical report, June 2022a. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Haohe_85_5.pdf`.

Junyan Liu, Zikai Zhou, Mengkai Sun, Kele Xu, Kun Qian, and Bian Hu. Seprotonet: Prototypical network with squeeze-and-excitation blocks for bioacoustic event detection. Technical report, DCASE2023 Challenge, June 2023.

Miao Liu, Jianqian Zhang, Lizhong Wang, Jiawei Peng, Chenguang Hu, Kaige Li, Jing Wang, and Qiuyue Ma. Bit srcb team ' s submission for dcase2022 task5 - few-shot bioacoustic event detection. Technical report, June 2022b. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Liu_43_5.pdf`.

Vincent Lostanlen, Justin Salamon, Mark Cartwright, Brian McFee, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters*, 26(1):39–43, 2018.

Regan D MacKinlay and Rachael C Shaw. A systematic review of animal personality in conservation science. *Conservation Biology*, 37(1):e13935, 2023.

Atefeh Mahdavi and Marco Carvalho. A survey on open set recognition. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 37–44. IEEE, 2021.

Aaquila Mariajohn. Bioacoustic few shot learning with class augmentation technical report. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Mariajohn_104_5.pdf`.

Hubert Markl. Stridulation in leaf-cutting ants. *Science*, 149(3690):1392–1393, 1965.

Killian Martin, Olivier Adam, Nicolas Obin, and Valérie Dufour. Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks. *bioRxiv*, 2022a.

Killian Martin, Olivier Adam, Nicolas Obin, and Valérie Dufour. Rookognise: Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks. *Ecological Informatics*, 72:101818, 2022b.

John Martinsson, Martin Willbo, Aleksis Pirinen, Olof Mogren, and Maria Sandsten. Few-shot bioacoustic event detection using a prototypical network ensemble with adaptive embedding functions. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Martinsson_78_5.pdf`.

Nicolas Mathevon, Isabelle Charrier, and Pierre Jouventin. Potential for individual recognition in acoustic signals: a comparative study of two gulls with different nesting patterns. *Comptes Rendus Biologies*, 326(3):329–337, 2003.

Kevin G McCracken and Frederick H Sheldon. Avian vocalizations and phylogenetic signal. *Proceedings of the National Academy of Sciences*, 94(8): 3833–3836, 1997.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. Sound event detection in the dcase 2017 challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(6):992–1006, 2019.

Marius Miron, David Robinson, Milad Alizadeh, Ellen Gilsenan-McMahon, Gagan Narula, Emmanuel Chemla, Maddie Cusimano, Felix Effenberger, Masato Hagiwara, Benjamin Hoffman, et al. What matters for bioacoustic encoding. *arXiv preprint arXiv:2508.11845*, 2025.

Veronica Morfi, Inês Nolasco, Vincent Lostanlen, Shubhr Singh, Ariana Strandburg-Peshkin, Lisa F Gill, Hanna Pamula, David Benvent, and Dan Stowell. Few-shot bioacoustic event detection: A new task at the dcase 2021 challenge. In *DCASE*, pages 145–149, 2021.

Ilyass Moummad, Romain Serizel, and Nicolas Farrugia. Supervised contrastive learning for pre-training bioacoustic few shot systems. Technical report, DCASE2023 Challenge, June 2023.

Ilyass Moummad, Romain Serizel, Emmanouil Benetos, and Nicolas Farrugia. Domain-invariant representation learning of bird sounds. *arXiv preprint arXiv:2409.08589*, 2024.

Tamara Münkemüller, Sébastien Lavergne, Bruno Bzeznik, Stéphane Dray, Thibaut Jombart, Katja Schiffers, and Wilfried Thuiller. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3(4):743–756, 2012.

Ran Nathan, Christopher T Monk, Robert Arlinghaus, Timo Adam, Josep Alós, Michael Assaf, Henrik Baktoft, Christine E Beardsworth, Michael G Bertram, Allert I Bijleveld, et al. Big-data approaches lead to an increased understanding of the ecology of animal movement. *Science*, 375(6582):eabg1780, 2022.

I. Nolasco, S. Singh, E. Vidała-Vila, E. Grout, J. Morford, M.G. Emmerson, F. H. Jensen, I. Kiskin, H. Whitehead, A. Strandburg-Peshkin, L. Gill, H. Pamuła, V. Lostanlen, V. Morfi, and D. Stowell. Few-shot bioacoustic event detection at the dcase 2022 challenge. In *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

Inês Nolasco and Dan Stowell. Rank-based loss for learning hierarchical representations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3623–3627. IEEE, 2022.

Inês Nolasco, Alessandro Terenzi, Stefania Cecchi, Simone Orcioni, Helen L Bear, and Emmanouil Benetos. Audio-based identification of beehive states. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8256–8260. IEEE, 2019.

Ines Nolasco, Burooj Ghani, Shubhr Singh, Ester Vidaña-Vila, Helen Whitehead, Emily Grout, Michael Emmerson, Ivan Kiskin, Frants Jensen, Joe Morford, Ariana Strandburg-Peshkin, Lisa Gill, Hanna Pamuła, Vincent Lostanlen, and Dan Stowell. Few-shot bioacoustic event detection at the DCASE 2023 challenge. In *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, pages 146–150, Tampere, Finland, September 2023a.

Ines Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana Strandburg-Peshkin, Ester Vidaña-Vila, Lisa Gill, Hanna Pamuła, Helen Whitehead, Ivan Kiskin, et al. Learning to detect an animal sound from five examples. *Ecological informatics*, 77:102258, 2023b.

Ines Nolasco, Ilyass Moummad, Dan Stowell, and Emmanouil Benetos. Acoustic identification of individual animals with hierarchical contrastive learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

Stephen Nowicki and Peter Marler. How do birds sing? *Music Perception*, 5 (4):391–426, 1988.

Stavros Ntalampiras and Ilyas Potamitis. Acoustic detection of unknown bird species and individuals. *CAAI Transactions on Intelligence Technology*, 6(3): 291–300, 2021.

Karan J Odom, Marcelo Araya-Salas, Janelle L Morano, Russell A Ligon, Gavin M Leighton, Conor C Taff, Anastasia H Dalziell, Alexis C Billings, Ryan R Germain, Michael Pardo, et al. Comparative bioacoustics: a roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews*, 96(4):1135–1159, 2021.

Vicente Palacios, Enrique Font, and Rafael Márquez. Iberian wolf howls: acoustic structure, individual variation, and a comparison with north american populations. *Journal of Mammalogy*, 88(3):606–613, 2007.

Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.

Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11814–11823, 2020.

Caroline Criado Perez. *Invisible women: the Sunday Times number one bestseller exposing the gender bias women face every day*. Random House, 2019.

Tereza Petrusková, Tomasz S Osiejuk, Pavel Linhart, and Adam Petrusek. Structure and complexity of perched and flight songs of the tree pipit (anthus trivialis). In *Annales Zoologici Fennici*, volume 45, pages 135–148. BioOne, 2008.

Tereza Petrusková, Iveta Pišvejcová, Anna Kinštová, Tomáš Brinke, and Adam Petrusek. Repertoire-based individual acoustic monitoring of a migratory passerine bird with complex song as an efficient tool for tracking territorial dynamics and annual return rates. *Methods in Ecology and Evolution*, 7(3): 274–284, 2016.

Lam Pham, Ian McLoughlin, Huy Phan, Ramaswamy Palaniappan, and Alfred Mertins. Deep feature embedding and hierarchical classification for audio scene classification. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

Maxine P Piggott and Andrea C Taylor. Remote collection of animal dna and its applications in conservation management and understanding the population biology of rare and cryptic species. *Wildlife Research*, 30(1):1–13, 2003.

Ilyas Potamitis, Todor Ganchev, and Nikos Fakotakis. Automatic acoustic identification of crickets and cicadas. *2007 9th International Symposium on Signal Processing and its Applications, ISSPA 2007, Proceedings*, pages 6–9, 2007. doi: 10.1109/ISSPA.2007.4555462.

Darren S Proppe and Emily Finch. Vocalizing during gaps in anthropogenic noise is an uncommon trait for enhancing communication in songbirds. *J Ecoacoustics*, 1, 2017.

Alexandra Průchová, Pavel Jaška, and Pavel Linhart. Cues to individual identity in songs of songbirds: testing general song characteristics in Chiffchaffs Phylloscopus collybita. *Journal of Ornithology*, 158(4):911–924, 2017. ISSN 21937206. doi: 10.1007/s10336-017-1455-6.

Jan A Randall. Drummers and stompers: vibrational communication in mammals. *The use of vibrations in communication: properties, mechanisms and function across taxa. Transworld, Kerala*, pages 99–120, 2010.

Denis Réale, Simon M Reader, Daniel Sol, Peter T McDougall, and Niels J Dingemanse. Integrating animal temperament within ecology and evolution. *Biological reviews*, 82(2):291–318, 2007.

Moises Rivera, Jacob A Edwards, Mark E Hauber, and Sarah MN Woolley. Machine learning and statistical classification of birdsong link vocal acoustic features with phylogeny. *Scientific reports*, 13(1):7076, 2023.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. *arXiv preprint arXiv:2411.07186*, 2024.

Holly Root-Gutteridge, Martin Bencsik, Manfred Chebli, Louise K Gentle, Christopher Terrell-Nield, Alexandra Bourit, and Richard W Yarnell. Identifying individual wild eastern grey wolves (canis lupus lycaon) using fundamental frequency and amplitude of howls. *Bioacoustics*, 23(1):55–66, 2014.

Ethan M Rudd, Lalit P Jain, Walter J Scheirer, and Terrance E Boult. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):762–768, 2017.

Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. (May), 2017. URL `http://arxiv.org/abs/1706.05098`.

Sougata Sadhukhan, Holly Root-Gutteridge, and Bilal Habib. Identifying unknown Indian wolves by their distinctive howls: its potential as a non-invasive survey method. *Scientific Reports*, 11(1):1–13, 2021. ISSN 20452322. doi: 10.1038/s41598-021-86718-w. URL `https://doi.org/10.1038/s41598-021-86718-w`.

Carl Safina. Beyond words: What animals think and feel. In *Farming, Food and Nature*, pages 75–85. Routledge, 2018.

Tom Schaul and Jürgen Schmidhuber. Metalearning. *Scholarpedia*, 5:4650, 2010.

Walter J. Scheirer, Anderson De Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2012.256.

Irena Schneiderová, Jan Pluháček, Simon Bearder, and Hanna Rosti. Bioacoustics reveals the uniqueness of an ex-situ tree hyrax population. *Journal of Zoo and Aquarium Research*, 12(1):42–48, 2024.

Stefan Scholl. Comparison of embedded spaces for deep learning classification. *arXiv preprint arXiv:2408.01767*, 2024.

Raphael Schwinger, Paria Vali Zadeh, Lukas Rauch, Mats Kurz, Tom Hauschild, Sam Lapp, and Sven Tomforde. Foundation models for bioacoustics–a comparative review. *arXiv preprint arXiv:2508.01277*, 2025.

Amanda Searby and Pierre Jouventin. Mother-lamb acoustic recognition in sheep: a frequency coding. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1526):1765–1771, 2003.

Sarab S Sethi, Nick S Jones, Ben D Fulcher, Lorenzo Picinali, Dena Jane Clink, Holger Klinck, C David L Orme, Peter H Wrege, and Robert M Ewers. Characterizing soundscapes across diverse ecosystems using a universal acoustic

feature set. *Proceedings of the National Academy of Sciences*, 117(29):17049–17055, 2020.

Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020.

Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011. ISSN 13845810. doi: 10.1007/s10618-010-0175-9.

Grace Smith-Vidaurre, Valeria Perez-Marrufo, and Timothy F Wright. Individual vocal signatures show reduced complexity following invasion. *Animal Behaviour*, 179:15–39, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Dan Stowell. *Computational Bioacoustic Scene Analysis*, pages 303–333. Springer International Publishing, Cham, 2018a. ISBN 978-3-319-63450-0. doi: 10.1007/978-3-319-63450-0_11. URL `https://doi.org/10.1007/978-3-319-63450-0_11`.

Dan Stowell. Computational bioacoustic scene analysis. In *Computational analysis of sound scenes and events*, pages 303–333. Springer, 2018b.

Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022.

Dan Stowell, Veronica Morfi, and Lisa F Gill. Individual identity in songbirds: signal representations and metric learning for locating the information in complex corvid calls. *arXiv preprint arXiv:1603.07236*, 2016.

Dan Stowell, Tereza Petrusková, Martin Šálek, and Pavel Linhart. Datasets for automatic acoustic identification of individual birds, October 2018. URL `https://doi.org/10.5281/zenodo.1413495`.

Dan Stowell, Tereza Petrusková, Martin Šálek, and Pavel Linhart. Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions. *Journal of the Royal Society Interface*, 16 (153), 2019. ISSN 17425662. doi: 10.1098/rsif.2018.0940.

Yizhou Tan, Lifan Xu, Chenyang Zhu, Shengchen Li, Haojun Ai, and Xi Shao. A new transductive framework for few-shot bioacoustic event detection task. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Tan_39_5.pdf`.

Jigang Tang, Xueyang Zhang, Tian Gao, Diyuan Liu, Jia Pan Xin Fang and, Qing Wang, Jun Du, Kele Xu, and Qinghua Pan. Few-shot embedding learning and event filtering for bioacoustic event detection. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Du_122_5.pdf`.

Anna M Taylor and David Reby. The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, 280(3):221–236, 2010.

Anna M. Taylor, Benjamin D. Charlton, and David Reby. *Vocal Production by Terrestrial Mammals: Source, Filter, and Function*, pages 229–259. Springer International Publishing, Cham, 2016. ISBN 978-3-319-27721-9. doi: 10.1007/978-3-319-27721-9_8. URL `https://doi.org/10.1007/978-3-319-27721-9_8`.

Andrew M.R. Terry, Tom M. Peake, and Peter K. McGregor. The role of vocal individuality in conservation. *Frontiers in Zoology*, 2:1–16, 2005. ISSN 17429994. doi: 10.1186/1742-9994-2-10.

Yu Teshima, Shoko Genda, Yota Aoki, Masahiro Fujisawa, Shizuko Hiryu, and Keisuke Fujii. Flight trajectory modeling reveals species-specific obstacle avoidance policies in echolocating bats. *bioRxiv*, pages 2025–06, 2025.

Michael D Thom and Jane L Hurst. Individual recognition by scent. In *Annales Zoologici Fennici*, pages 765–787. JSTOR, 2004.

Joseph A Tobias and Alex L Pigot. Integrating behaviour and ecology into global biodiversity conservation strategies. *Philosophical Transactions of the Royal Society B*, 374(1781):20190012, 2019.

Rose Trappes, Alkistis Elliott-Graves, and Marie I Kaiser. Studying individuality in behavioral ecology: Overcoming epistemic challenges. *Perspectives on Science*, pages 1–50, 2025.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. Hear 2021: Holistic evaluation of audio representations. *arXiv preprint arXiv:2203.03022*, 2022.

Bart Van Merriënboer, Jenny Hamer, Vincent Dumoulin, Eleni Triantafillou, and Tom Denton. Birds, bats and beyond: Evaluating generalization in bioacoustics models. *Frontiers in Bird Science*, 3:1369756, 2024.

Maxime Vidal, Nathan Wolf, Beth Rosenberg, Bradley P Harris, and Alexander Mathis. Perspectives on individual animal identification from biology and computer vision. *Integrative and comparative biology*, 61(3):900–916, 2021.

Manuel Vieira, Paulo J Fonseca, M Amorim, and Carlos JC Teixeira. Call recognition and individual identification of fish vocalizations based on automatic speech recognition: an example with the lusitanian toadfish. *The Journal of the Acoustical Society of America*, 138(6):3941–3950, 2015a.

Manuel Vieira, Paulo J. Fonseca, M. Clara P. Amorim, and Carlos J. C. Teixeira. Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example with the Lusitanian toadfish. *The Journal of the Acoustical Society of America*, 138(6):3941–3950, 2015b. ISSN 0001-4966. doi: 10.1121/1.4936858. URL `http://dx.doi.org/10.1121/1.4936858`.

Tuomas Virtanen, Mark D Plumbley, and Dan Ellis. *Computational analysis of sound scenes and events*, volume 9. Springer, 2018.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020a. doi: 10.1145/3386252.

Yu Wang, Justin Salamon, Mark Cartwright, Nicholas J. Bryan, and Juan Pablo Bello. Few-shot drum transcription in polyphonic music. *CoRR*, abs/2008.02791, 2020b. URL `https://arxiv.org/abs/2008.02791`.

Malachi Whitford and A Peter Klimley. An overview of behavioral, physiological, and environmental sensors used in animal biotelemetry and biologging studies. *Animal Biotelemetry*, 7(1):1–24, 2019.

Matthew Wijers, Paul Trethowan, Byron Du Preez, Simon Chamaillé-Jammes, Andrew J Loveridge, David W Macdonald, and Andrew Markham. Vocal discrimination of african lions and its potential for collar-free tracking. *Bioacoustics*, 30(5):575–593, 2021.

Kevin Wilkinghoff and Alessia Cornaggia-Urrigshardt. Few-shot bioacoustic event detection. Technical report, DCASE2023 Challenge, June 2023.

Martin Willbo, John Martinsson, Aleksis Pirinen, and Olof Mogren. Wide resnet models for few-shot sound event detection. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Willbo_53_5.pdf`.

Piper Wolters, Chris Daw, Brian Hutchinson, and Lauren Phillips. Proposal-based few-shot sound event detection for speech and environmental sounds with perceivers. *arXiv preprint arXiv:2107.13616*, 2021.

Xiaoxiao Wu and Yanhua Long. Few-shot continual learning for bioacoustic event detection. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Wu_4_5.pdf`.

Yong Xu, Qiang Huang, Wenwu Wang, and Mark D. Plumbley. Hierarchical learning for DNN-based acoustic scene classification. (September), 2016. URL `http://arxiv.org/abs/1607.03682`.

Genwei Yan, Ruoyu Wang, Liang Zou, Jun Du, Qing Wang, Tian Gao, and Xin Fang. Multi-task frame level system for few-shot bioacoustic event detection. Technical report, DCASE2023 Challenge, June 2023.

Dongchao Yang, Yuexian Zou, Fan Cui, and Yujun Wang. Improved prototypical network with data augmentation. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Zou_36_5.pdf`.

Sophia Yin and Brenda McCowan. Barking in domestic dogs: Context specificity and individual identification. *Animal Behaviour*, 68(2):343–355, 2004. ISSN 00033472. doi: 10.1016/j.anbehav.2003.07.016.

Jessica L. Yorzinski. Peafowl antipredator calls encode information about signalers. *The Journal of the Acoustical Society of America*, 135(2):942–952, 2014. ISSN 0001-4966. doi: 10.1121/1.4861340.

Jessica L Yorzinski. The cognitive basis of individual recognition. *Current Opinion in Behavioral Sciences*, 16:53–57, 2017.

Liwen You, Erika Pelaez Coyotl, Suren Gunturu, and Maarten Van Segbroeck. Transformer-based bioacoustic sound event detection on few-shot learning tasks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, et al. Marble: Music audio representation benchmark for universal evaluation. *Advances in Neural Information Processing Systems*, 36:39626–39647, 2023.

Bartlomiej Zgorzynski and Mateusz Matuszewski. Siamese network for few-shot bioacoustic event detection. Technical report, June 2022. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Zgorzynski_55_5.pdf`.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, 2022a.

Tianyang Zhang, Yuyang Wang, and Ying Wang. A meta-learning framework for few-shot sound event detection. Technical report, June 2022b. URL `https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Zhang_6_5.pdf`.

Yu Zhang and Qiang Yang. A Survey on Multi-Task Learning. 2017. URL `http://arxiv.org/abs/1707.08114`.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10875, 2021.