MARINHA DO BRASIL

INSTITUTO DE ESTUDOS DO MAR ALMIRANTE PAULO MOREIRA

UNIVERSIDADE FEDERAL FLUMINENSE

PROGRAMA ASSOCIADO DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA MARINHA

VIVIANE RODRIGUES BARROSO

DETECÇÃO E CLASSIFICAÇÃO DE SONS DE PEIXES MARINHOS UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

ARRAIAL DO CABO / RJ
2026

VIVIANE RODRIGUES BARROSO

DETECÇÃO E CLASSIFICAÇÃO DE SONS DE PEIXES MARINHOS UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Tese de doutorado apresentada ao Instituto de Estudos do Mar Almirante Paulo Moreira e à Universidade Federal Fluminense, como requisito parcial para a obtenção do grau de Doutor em Biotecnologia Marinha.

Orientador: Prof. Dr. Carlos E.L. Ferreira
Coorientador: Dr. Fabio Contrera Xavier

ARRAIAL DO CABO / RJ
2026

FICHA CATALOGRÁFICA

Biblioteca do IEAPM

B277d Barroso, Viviane R.

Detecção e classificação de sons de peixes marinhos utilizando técnicas de inteligência artificial / [por] Viviane Rodrigues Barroso.-- 2026.

70 f.: il.; 30 cm.

Orientador: Carlos E. L. Ferreira

Coorientador: Fabio Contrera Xavier

Tese (Doutorado) - Instituto de Estudos do Mar Almirante Paulo Moreira e Universidade Federal Fluminense - IEAPM/UFF, Programa Associado de Pós-Graduação em Biotecnologia Marinha, Arraial do Cabo, 2026.

1. Aprendizado de máquina 2. Redes neurais 3. Embeddings multimodais 4. Classificação de sons 5. Biodiversidade I. Ferreira, Carlos E. L. II. Instituto de Estudos do Mar Almirante Paulo Moreira. Programa Associado de Pós-Graduação em Biotecnologia Marinha. III. Universidade Federal Fluminense IV. Título.

CDD: Ed. 22 -- 660.6

VIVIANE RODRIGUES BARROSO

DETECÇÃO E CLASSIFICAÇÃO DE SONS DE PEIXES MARINHOS UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Tese de doutorado apresentada ao Instituto de Estudos do Mar Almirante Paulo Moreira e à Universidade Federal Fluminense, como requisito parcial para a obtenção do grau de Doutor em Biotecnologia Marinha

COMISSÃO JULGADORA:

____
Prof. Dr. Carlos Eduardo Leite Ferreira
Universidade Federal Fluminense
Professor orientador - Presidente da banca examinadora

____
Prof. Dra. Renata Santoro Sousa-Lima
Universidade Federal do Rio Grande do Norte

____
Prof. Dr. Rogério de Moraes Calazan
Instituto de Estudos do Mar Almirante Paulo Moreira

____
Prof. Dr. Bernardo Antonio Perez da Gama
Universidade Federal Fluminense

____
Prof. Dr. Eduardo Barros Fagundes Netto
Instituto de Estudos do Mar Almirante Paulo Moreira

Arraial Do Cabo/RJ, 06 de fevereiro de 2026.

"O objetivo da computação é a compreensão,
não os números."

(Richard Hamming)

AGRADECIMENTOS

Agradeço a todos que contribuíram para a realização deste trabalho.

Em primeiro lugar, agradeço à minha família e amigos pelo apoio e incentivo. Agradeço especialmente à minha mãe Sônia pelo amor e dedicação e ao meu marido Fabricio pelo incentivo e por sempre acreditar nos meus sonhos.

Agradeço aos meus gatos Carmilla, Golias, Yumi, Clotilde e Joaquim pela companhia!

Às minhas amigas Bessie Ribeiro, Débora Persich e Juliana Santos pelo apoio e incentivos.

Aos amigos da Bioacústica e do Projeto ECOMAC: aos amigos Ubirajara Melo e Murilo Minello pelas colaborações. Agradeço especialmente às minhas amigas Aléxia Lessa e Ana Luz pelo apoio e pelas colaborações.

Ao meu orientador Carlos Ferreira por acreditar nesse trabalho e pelos constantes incentivos.

Ao meu coorientador Fabio Contrera pelas contribuições, por todo incentivo, apoio, encorajamento e pelos conselhos.

Aos membros da banca examinadora por terem aceitado o convite para a avaliação deste trabalho e pelas valiosas sugestões.

À Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento desta pesquisa.

RESUMO

O estudo da bioacústica de peixes por meio do Monitoramento Acústico Passivo (MAP) é essencial para a compreensão da biodiversidade marinha e a avaliação dos impactos de estressores antropogênicos sobre os ecossistemas aquáticos. No entanto, essa área ainda apresenta lacunas decorrentes da complexidade das paisagens sonoras subaquáticas e da escassez de conjuntos de dados rotulados, validados e específicos por espécie. Esta tese explora a evolução e a aplicação da inteligência artificial (IA) na automação da detecção e classificação de sons de peixes. A fundamentação desta pesquisa estruturou-se a partir de uma análise crítica do estado da arte, que evidenciou a necessidade de métodos que superem a limitação da escassez de dados rotulados e aumentem a interpretação e transparência dos modelos. Diante disso, estabeleceu-se um *framework* de aprendizado supervisionado para a classificação de sons pulsados de peixes em um ambiente de recife rochoso subtropical. Com a incorporação de técnicas de aumento de dados, os modelos, em particular o Perceptron Multicamadas, alcançaram uma acurácia de 98,1%. A inovação desta etapa reside no uso de Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI), que ampliou a transparência e a interpretabilidade dos modelos, além da identificação inédita de características acústicas determinantes, conferindo uma base de interpretação biológica às decisões automatizadas dos modelos. Como uma alternativa para superar a lacuna de escassez de dados rotulados, a pesquisa avançou para o uso do *Contrastive Language–Audio Pre training* (CLAP). Por meio de recuperação cruzada entre modalidades e classificação *zero-shot*, essa abordagem multimodal demonstrou uma capacidade robusta de reconhecer padrões acústicos sem a necessidade de grandes conjuntos de dados para treinamento. Destaca-se o desempenho expressivo para a família Sciaenidae, que respondeu por mais de 60% das classificações corretas. Em conjunto, esses resultados indicam que, embora modelos supervisionados apresentem alta precisão na classificação desses sons, embeddings multimodais e técnicas de aprendizado *zero-shot* oferecem maior flexibilidade para um monitoramento acústico marinho escalável e taxonomicamente informado.

Palavras-chaves: aprendizado de máquina, redes neurais, *embeddings* multimodais, classificação de sons, biodiversidade.

ABSTRACT

The study of fish bioacoustics through passive acoustic monitoring (PAM) is essential to understanding marine biodiversity and assessing the impact of human activities on aquatic ecosystems. However, gaps remain in this area due to the complexity of underwater soundscapes and the scarcity of labeled and validated species-specific datasets. This thesis explores the evolution and application of artificial intelligence (AI) for automating fish sound detection and classification. Building on a critical review of the state-of-the-art, this study identified the need for methods capable of overcoming labeled data scarcity and enhancing model transparency and interpretability. Thus, a supervised learning framework was developed to classify pulsed fish sounds in a subtropical rocky reef environment. With the incorporation of data augmentation techniques, the models, particularly the multilayer perceptron, achieved an accuracy of 98.1%. This stage innovates by using Explainable Artificial Intelligence (XAI), which increases the models' transparency and interpretability, as well as the identification of determining acoustic characteristics. This provides a biological basis for interpreting the models' automated decisions. To address the scarcity of labeled data, the research evolved toward the use of Contrastive Language–Audio Pretraining (CLAP). Through cross-modal retrieval and zero-shot classification, this multimodal approach demonstrated a robust capability to recognize acoustic patterns without the need for large training datasets. Notably, the model showed significant performance for the Sciaenidae family, which accounted for over 60% of correct classifications. Taken together, these results indicate that while supervised models yield high precision in sound classification, multimodal embeddings and zero-shot learning techniques offer greater flexibility for scalable and taxonomically informed marine acoustic monitoring

Keyowrds: machine learning, neural networks, multimodal embeddings, sound classification, biodiversity.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BERT	*Bidirectional Encoder Representations from Transformers* (*encoder* de texto)
BIOCOM	Projeto de Construção de Métodos de Sinalização com Características de Ruídos Bioacústicos para Comunicação Submarina.
CLAP	*Contrastive Language-Audio Pretraining* (modelo CLAP)
CLIP	*Contrastive Language-Image Pretraining* (modelo CLIP)
CNN	*Convolutional Neural Network* (Rede Neural Convolucional)
DL	*Deep Learning* (Aprendizado Profundo)
DOSITS	*Discovery of Sound in the Sea* (website)
DT	*Decision Trees* (Árvores de Decisão)
EDA	*Exploratory Data Analysis* (Análise Exploratória de Dados)
EOV	*Essential Ocean Variable* (Variável Oceânica Essencial)
GSO	*Graduate School of Oceanography (Fish Sounds Collection)* - Base de dados de sons de peixes
HTSAT	*Hierarchical Token-Semantic Audio Transformer* (*encoder* de áudio)
IA	Inteligência Artificial
IPI	*Interpulse interval* (intervalo inter-pulso)
MFCC	*Mel Frequency Cepstral Coefficients* (Coeficientes Cepstrais de Frequência Mel)
ML	*Machine Learning* (Aprendizado de Máquina)
MLP	*Multilayer Perceptron* (Perceptron Multicamadas)
MMR	*Mean Reciprocal Rank* (métrica de performance para sistemas de recuperação)
NB	*Naive Bayes* (classificador Naive Bayes)
NN	*Neural Network* (Rede Neural)
PAM	*Passive Acoustic Monitoring* (Monitoramento Acústico Passivo - MAP)
RF	*Random Forest* (Floresta Aleatória)

RoBERTa	*Robustly optimized BERT approach* (*encoder* de texto)
SHAP	*SHapley Additive exPlanations* (Explicações Aditivas de Shapley)
SotA	*State-of-the-Art* (Estado da arte)
WoRMS	*World Register of Marine Species* (Base de dados para organismos marinhos)
XAI	*Explainable Artificial Intelligence* (Inteligência Artificial Explicável)

SUMÁRIO

1. INTRODUÇÃO GERAL

Os ecossistemas marinhos têm sido impactados por múltiplos estressores antrópicos, tais como sobrepesca, degradação de habitats, acidificação, mudanças climáticas globais e poluição sonora (DUARTE et al., 2021). Esses fatores podem levar à perda de biodiversidade e afetar processos críticos de resiliência e funcionamento desses sistemas (OLIVER et al., 2015). A biodiversidade de peixes é particularmente importante, uma vez que esses organismos são componentes-chave, desempenhando funções ecológicas críticas, como a regulação da estrutura trófica e contribuindo para a resiliência dos sistemas (HOLMLUND; HAMMER, 1999). Dessa forma, torna-se essencial o desenvolvimento de estratégias de monitoramento que permitam acompanhar de forma eficiente essas mudanças nos padrões de biodiversidade frente aos cenários atuais de múltiplos estressores.

A avaliação e o monitoramento da biodiversidade de peixes são tradicionalmente baseadas em métodos visuais não destrutivos (BURGI et al., 2025) e/ou destrutivos (eg., *fishing gears*) (MUNDAY; WILSON, 1997; ROBERTSON; SMITH-VANIS, 2010). Esses métodos apresentam limitações relacionadas à cobertura espaço-temporal e a aspectos operacionais. Por outro lado, o monitoramento acústico passivo (MAP) emerge como um método complementar eficiente, permitindo a aquisição contínua, remota e não invasiva de grandes volumes de dados acústicos subaquáticos (GANNON, 2008). Nesse contexto, o MAP representa uma ferramenta poderosa para o estudo de peixes, considerando que muitas espécies de peixes produzem sons de comunicação (LOOBY et al., 2022; RICE et al., 2022) associados a diferentes contextos comportamentais (KASUMYAN, 2009; LADICH, 2019), que contribuem de forma significativa para as paisagens acústicas marinhas. Esses sons apresentam padrões espectrais e temporais característicos, frequentemente pulsados e repetitivos, o que viabiliza a sua detecção e, em determinados casos, a identificação em nível taxonômico (WALL et al., 2013; PUTLAND et al., 2018). Contudo, a aplicação do MAP em larga escala enfrenta desafios que residem tanto na natureza dos dados quanto na lacuna de conhecimento bioacústico sobre essas espécies.

Uma das principais limitações reside no fato de que ainda não se sabe exatamente quantas e quais espécies produzem sons (PARMENTIER et al., 2021).

Entre aproximadamente 35.000 espécies conhecidas (FROESE; PAULY, 2025), apenas cerca de 1.400 espécies, pertencentes a 175 famílias, foram documentadas quanto à capacidade de produzir sons (RICE et al., 2022). Além disso, entre as espécies registradas, poucas têm seus repertórios acústicos completamente caracterizados (PARSONS et al., 2022). Isso significa que possivelmente existe um número muito maior de espécies capazes de produzir sons de comunicação e que, para uma mesma espécie, os repertórios vocais podem ser mais diversos do que os atualmente descritos. Portanto, muitos sons detectados nas gravações não podem ser prontamente atribuídos a uma espécie conhecida, o que dificulta a resolução taxonômica.

As gravações de longo período de MAP resultam em conjuntos de dados massivos e heterogêneos, caracterizados por grande volume, diversidade de fontes e baixa relação sinal-ruído, nos quais os sinais de interesse estão frequentemente imersos em ruído ambiente (HILDEBRAND, 2009). Esses fatores tornam a análise manual inviável e evidenciam a necessidade de métodos computacionais robustos, escaláveis e interpretáveis para a detecção e classificação desses sons. Nesse contexto, avanços recentes em Inteligência Artificial (IA) têm possibilitado a construção de métodos computacionais capazes de extrair representações acústicas discriminativas e automatizar processos de detecção e classificação (BIANCO et al. 2019).

Diante desse cenário, a presente tese aborda a detecção e classificação de sons de peixes a partir de uma perspectiva computacional avançada. O trabalho explora o uso de aprendizado de máquina aliado a técnicas de interpretabilidade de modelos (*Explainable Artificial Intelligence* – XAI), bem como estratégias de aprendizado contrastivo utilizando representações vetoriais multimodais de áudio e texto. Ao integrar a classificação supervisionada de sons desconhecidos de peixes recifais, aliada à interpretabilidade e transparência dos modelos e ao investigar o uso de *embeddings* multimodais em contextos de classificação zero-shot, esta pesquisa propõe uma estratégia para ampliar a generalização semântica dos modelos de classificação, o que representa uma fronteira metodológica promissora para aplicações na bioacústica de peixes. Uma vez que alguns desafios nesse campo sejam superados, a Inteligência Artificial (IA) aplicada ao MAP pode representar o futuro do monitoramento de biodiversidade de peixes, complementando outras técnicas mais usuais e não destrutivas (ou invasivas).

1.1. OBJETIVOS

1.1.1. Objetivo geral

Desenvolver e avaliar métodos computacionais para classificação de sons de peixes, integrando técnicas de interpretabilidade de modelos e o uso de *embeddings* multimodais de áudio e texto.

1.1.2. Objetivos específicos

- Realizar uma revisão do estado da arte sobre detecção, classificação e identificação de sons de peixes por meio de técnicas de *machine learning* e *deep learning*.
- Desenvolver um classificador supervisionado para sons de peixes recifais, incorporando estratégias de aumento de dados e métodos de interpretabilidade de modelos (XAI).
- Avaliar a aplicação de *embeddings* multimodais (áudio e texto) para classificação *zero-shot* de sons de peixes.

1.2. ORGANIZAÇÃO DO DOCUMENTO

Com o objetivo de desenvolver e avaliar métodos computacionais para classificação de sons de peixes utilizando diferentes técnicas de inteligência artificial, esta tese foi estruturada em 3 capítulos. O Capítulo 1, intitulado "*Applications of machine learning to identify and characterize the sounds produced by fish*", consiste em uma revisão bibliográfica sobre a utilização de algoritmos de *machine learning* e *deep learning* para detecção, classificação e identificação de sons de peixes, incluindo uma discussão sobre os desafios e alguns pontos para guiar estudos futuros. Essa revisão foi publicada no *Journal of Marine Science* em 2023. O Capítulo 2, intitulado "*Optimal feature selection and model explanation for reef fish sound classification*", trata do desenvolvimento de um classificador de sons desconhecidos de peixes recifais, utilizando quatro algoritmos de aprendizado supervisionados e *Explainable Artificial Intelligence* (XAI) para explicar o modelo de classificação e a influência das características dos sons na previsão de cada classe de som. Este artigo foi publicado em um *Special issue* intitulado "*Acoustic monitoring for tropical ecology and conservation*" da revista *Philosophical Transactions of the Royal Society B* em 2025. Por fim, o Capítulo 3, intitulado "*Exploring the feasibility of multimodal embeddings for fish sound classification*", trata do uso de *embeddings* multimodais (áudio e texto) para a classificação *zero-shot* de sons produzidos por peixes. Além disso, este documento inclui uma introdução geral (apresentada na seção anterior), que integra os temas abordados nos três capítulos, bem como uma seção de considerações finais. Ao final do documento, também são disponibilizados materiais suplementares. No Apêndice 1 encontra-se o material suplementar para o artigo apresentado no Capítulo 2 e, no Apêndice 2, o material suplementar para o artigo apresentado no Capítulo 3. No Apêndice 3 encontram-se os artigos e capítulos de livros publicados. Por fim, no Apêndice 4, encontram-se as revisões de artigos realizadas para periódicos durante o período do doutorado. Para auxiliar na compreensão da terminologia técnica utilizada, esta tese inclui um glossário posicionado antes dos apêndices.

1.3. REFERÊNCIAS BIBLIOGRÁFICAS

BIANCO, M. J.; GERSTOFT, P.; TRAER, J.; OZANICH, E.; ROCH, M. A.; GANNOT, S.; DELEDALLE, C. A. Machine learning in acoustics: theory and applications. *The Journal of the Acoustical Society of America*, v. 146, p. 3590–3628, 2019.

BÜRGI, K.; BOUVEYRON, C.; LINGRAND, D.; DERIJARD, B.; PRECIOSO, F.; SABOURAULT, C. Towards a fully automated underwater census for fish assemblages in the Mediterranean Sea. *Ecological Informatics*, v. 85, p. 102959, 2025.

DUARTE, C. M.; CHAPUIS, L.; COLLIN, S. P.; COSTA, D. P.; EGUILUZ, V. M.; ERBE, C.; GORDON, T. A.; HALPERN, B. S.; HAVLIK, M. N. The soundscape of the Anthropocene ocean. *Science*, v. 371, p. eaba4658, 2021.

FROESE, R.; PAULY, D. (ed.). FishBase. World Wide Web electronic publication. Disponível em: <www.fishbase.org>. Versão: 11/2025.

GANNON, D. P. Passive acoustic techniques in fisheries science: a review and prospectus. *Transactions of the American Fisheries Society*, v. 137, n. 2, p. 638–656, 2008.

HILDEBRAND, J. A. Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series*, v. 395, p. 5–20, 2009.

HOLMLUND, C. M.; HAMMER, M. Ecosystem services generated by fish populations. *Ecological Economics*, v. 29, n. 2, p. 253–268, 1999.

KASUMYAN, A. O. Acoustic signaling in fish. *Journal of Ichthyology*, v. 49, p. 963–1020, 2009.

LADICH, F. Ecology of sound communication in fishes. *Fish and Fisheries*, v. 20, n. 3, p. 552–563, 2019.

LOOBY, A.; COX, K.; BRAVO, S.; ROUNTREE, R.; JUANES, F.; REYNOLDS, L. K.; MARTIN, C. W. A quantitative inventory of global soniferous fish diversity. *Reviews in Fish Biology and Fisheries*, v. 32, n. 2, p. 581–595, 2022.

MUNDAY, P. L.; WILSON, S. K. Comparative efficacy of clove oil and other chemicals in anaesthetization of *Pomacentrus amboinensis*, a coral reef fish. *Journal of Fish Biology*, v. 51, n. 5, p. 931–938, 1997.

OLIVER, T. H.; HEARD, M. S.; ISAAC, N. J.; ROY, D. B.; PROCTER, D.; EIGENBROD, F.; FRECKLETON, R.; HECTOR, A.; ORME, C. D. L.; PETCHEY, O. L.; PROENÇA, V. Biodiversity and resilience of ecosystem functions. *Trends in Ecology & Evolution*, v. 30, n. 11, p. 673–684, 2015.

PARMENTIER, E.; BERTUCCI, F.; BOLGAN, M.; LECCHINI, D. How many fish could be vocal? An estimation from a coral reef (Moorea Island). *Belgian Journal of Zoology*, v. 151, p. 82, 2021.

PARSONS, M. J.; LIN, T. H.; MOONEY, T. A.; ERBE, C.; JUANES, F.; LAMMERS, M.; LI, S. et al. Sounding the call for a global library of underwater biological sounds. *Frontiers in Ecology and Evolution*, v. 10, p. 39, 2022.

PUTLAND, R. L.; MACKIEWICZ, A. G.; MENSINGER, A. F. Localizing individual soniferous fish using passive acoustic monitoring. *Ecological Informatics*, v. 48, p. 60–68, 2018.

RICE, A. N.; FARINA, S. C.; MAKOWSKI, A. J.; KAATZ, I. M.; LOBEL, P. S.; BEMIS, W. E.; BASS, A. H. Evolutionary patterns in sound production across fishes. *Ichthyology & Herpetology*, v. 110, n. 1, p. 1–12, 2022.

ROBERTSON, D. R.; SMITH-VANIZ, W. F. Use of clove oil in collecting coral reef fishes for research. *Marine Ecology Progress Series*, v. 401, p. 295–302, 2010.

WALL, C. C.; SIMARD, P.; LEMBKE, C.; MANN, D. A. Large-scale passive acoustic monitoring of fish sound production on the West Florida Shelf. *Marine Ecology Progress Series*, v. 484, p. 173–188, 2013.

2. CAPÍTULO 1 - APPLICATIONS OF MACHINE LEARNING TO IDENTIFY AND CHARACTERIZE THE SOUNDS PRODUCED BY FISH

Published in the ICES Journal of Marine Science, v.80, n.7 (September, 2023). DOI: https://doi.org/10.1093/icesjms/fsad126.

ABSTRACT

Aquatic ecosystems are constantly changing due to anthropic stressors, which can lead to biodiversity loss. Ocean sound is considered an essential ocean variable, with the potential to improve our understanding of its impact on marine life. Fish produce a variety of sounds and their choruses often dominate underwater soundscapes. These sounds have been used to assess communication, behaviour, spawning location, and biodiversity. Artificial intelligence can provide a robust solution to detect and classify fish sounds. However, the main challenge in applying artificial intelligence to recognize fish sounds is the lack of validated sound data for individual species. This review provides an overview of recent publications on the use of machine learning, including deep learning, for fish sound detection, classification, and identification. Key challenges and limitations are discussed, and some points to guide future studies are also provided.

Keywords: artificial intelligence, bioacoustics, fish calls, passive acoustic monitoring, sound classification

1. INTRODUCTION

Aquatic ecosystems have been under constant changes due to anthropic stressors, such as overfishing, habitat degradation, and climate change, including ocean acidification (HALPERN et al., 2008). Noise pollution is also a threat to marine ecosystems and has increased significantly in recent years (DUARTE et al., 2021). These changes can lead to biodiversity loss and affect critical processes of their resilience and functioning (OLIVER et al., 2015). Fish biodiversity is of particular concern, as they are key components with critical ecological functions and help to regulate the trophic structure, which contributes to system resilience (HOLMLUND; HAMMER, 1999).

Different destructive methods are employed to determine fish composition and abundance, including many fishing gears and ichthyocides (e.g. rotenone, quinaldine, and clove oil) (MUNDAY; WILSON, 1997; ROBERTSON; SMITH-VANIS, 2010). However, non-destructive methods, including underwater visual census (UVC), remote videos (baited remote underwater videos—BRUVs), and diver-operated stereo-video systems (DOVs), are more widely used as fish monitoring tools (MURPHY; JENKINS, 2010; OSGOOD et al., 2019; HELLMRICH et al., 2023). While all of these methods present limitations for different specific targets and many robust comparisons are available (WILLIS, 2001; ACKERMAN; BELLWOOD, 2002; HARVEY et al., 2004; WATSON et al., 2005; GALAIDUK et al., 2017; FRENCH et al., 2021; ROLIM et al., 2022), successful long-term monitoring programmes use them as complementary methods (MALLET et al., 2014; CHEAL et al., 2021). Environmental DNA (eDNA) appears as a prospective non-destructive, easily replicable monitoring tool for marine biodiversity (LAMY et al., 2021; JUHEL et al., 2022); however, it is still limited by many constraints, such as the influence of environmental conditions on its concentration in water samples (LACOURSIÈRE-ROUSSEL et al., 2016).

Passive acoustic approaches, on the other hand, have the advantage of being remote, low-cost, and providing wider spatiotemporal coverage (GANNON, 2008). While human-made noise in the ocean is a major stressor, the use of sound through passive techniques can be a tool to assess and monitor biodiversity for conservation purposes. Since sound travels efficiently underwater, and because over a thousand fish species use sound to communicate (LOOBY et al., 2022a; RICE et al., 2022), passive acoustic monitoring (PAM) is a feasible alternative to traditional methods for assessing fish communities. It is worth mentioning that ocean sound is recognized as an essential ocean variable (EOV), with the potential to improve the understanding of soundscapes, its impacts on marine organisms, and its use in assessing biodiversity and ecosystem health (MIKSIS-OLDS et al., 2018).

Fish communication sounds can provide information on their behaviour, distribution, reproduction, and spawning (LOBEL, 2002; MANN, 2012; LADICH, 2019). Furthermore, these sounds can be used as proxies to infer taxonomic diversity and relative abundance (DESIDERÀ et al., 2019). Nonetheless, manual analysis of acoustic data to identify fish sounds is a challenging and time-consuming task due to the large amount of data collected in a complex and dynamic environment. Moreover, acoustic signals produced by fish vary in temporal and spectral characteristics (AMORIM, 2006), and these signals are embedded in ambient noise (HILDEBRAND, 2009). Thus, there is a need for reliable automatic detection, identification, and classification of fish sounds. Recent advances in artificial intelligence (AI) provide fast and robust techniques for these tasks with promising results (BIANCO et al., 2019). Once some challenges have been overcome, AI applied to PAM data may represent the future of fish monitoring, complementing more usual techniques.

Due to the increasing attention that passive acoustics has received, some literature reviews on this topic have emerged in the last few years (MELLINGER et al., 2007; VAN PARIJS et al., 2009; LINDSETH; LOBEL, 2018; MOONEY et al., 2020). However, none of them focused on the use of artificial intelligence for fish sound identification and characterization. Motivated by the significant amount of recent work using AI to detect, classify, and identify fish sounds, here we review the machine learning and deep learning techniques for the assessment of sounds produced by fish. This review does not focus on all applications of PAM, but on how machine learning, including deep learning, can aid the characterization of fish sounds and their use to provide estimates of fish acoustic diversity.

2. FISH SOUNDS

Fish are an important component of aquatic biophony, as they produce a wide variety of sounds. Nearly all fish species can detect sounds, and > 900 species from 133 families are known to produce communication sounds (LOOBY et al., 2022a), out of ~35000 documented fish species (FROESE; PAULY, 2022).

Sound production involves different structures that contribute to fish biophony in different ways. The most common mechanism is the vibration of the swim bladder by the contraction of sonic muscles, which produces low-frequency and often harmonic sounds (LADICH, 2014). Another common mechanism is known as stridulation, in which bony parts such as teeth, spine, and fin rays rub against others (TAVOLGA, 1971). This mechanism produces broadband pulsed sounds with frequencies on the order of a few thousand hertz (Hz) (LADICH, 2014). Recently, Rice et al. (2022) have mapped sound production in the Actinopterygii clade and, based on morphological characteristics, found 60 families with muscles attached to the swim bladder, 39 families that use the movement of bony plates to produce sound (stridulation mechanism), and 18 of these families with both mechanisms.

The diversity of sound production mechanisms, which is the greatest among vertebrates (LADICH, 2014), contributes to the production of different sound types, with significant variation between species (GANNON, 2008). These sound types are commonly named after their temporal and spectral characteristics, such as “pulse series”, “fast pulse train”, and “long tonal call”, among others (DESIDERÀ et al., 2019). According to Amorim (2006), variations in the temporal characteristics of fish sounds are the most common, with a diversity of tonal sound patterns and pulses. Thus, different sound types are not only important to detect different behaviours, but are also fundamental acoustic features for detection, identification, and classification algorithms. Besides, fish sounds are often repetitive and distinct, which makes species identification feasible (WALL et al., 2013; PUTLAND et al., 2018).

3. CLASSICAL APPROACHES

Classical approaches consist of analysing audio files generated by in situ recordings to detect fish calls. PAM in the marine environment employs different platforms equipped with hydrophones, such as mobile platforms or fixed autonomous recorders (MERCHANT et al., 2015). The basic technique consists of extracting acoustic metrics by inspecting spectrograms (MANN et al., 2008; LIN et al., 2018), which are computed using Fourier transforms to determine the presence of target species and to identify biological sounds at a specific level. Acoustic metrics of fish communication signals can include frequency, fundamental frequency, peak frequency, sound pressure level (SPL), and duration, among others (FINE; LENHARDT, 1983; TELLECHEA et al., 2010; DI IORIO et al., 2018). As fish often produce sequences of pulses (AMORIM, 2006; GANNON, 2008), other acoustic features related to the temporal structure of signals and rhythm can be included. Examples of these features are call duration, call rate, pulse duration, number of pulses, pulse rate (GANNON, 2008; RUPPÉ et al., 2015; DI IORIO et al., 2018), repetition rate, inter-pulse interval (IPI) (TELLECHEA et al., 2010; CARRIÇO et al., 2020a), and inter-onset interval (IOI) (BURCHARDT et al., 2021).

PAM has been widely applied to fish population studies in a variety of contexts and environments. One example is the substantial work that has been done to locate and monitor spawning sites and to characterize sounds produced during spawning periods (ROWELL et al., 2017; JUBLIER et al., 2020; WILSON et al., 2022). However, few studies have focused exclusively on quantifying biodiversity metrics (e.g. abundance, species richness, density, or biomass). This can be explained by the low correlation between bioacoustic diversity and taxonomic diversity in the marine environment (DESIDERÀ et al., 2019). In addition, there are limitations in relating sound levels to measurements of abundance and biomass (ROWELL et al., 2017). Nevertheless, some studies have used passive acoustics to estimate relative abundances (ROWELL et al., 2012; SCHÄRER et al., 2012; DESIDERÀ et al., 2019). Some authors suggest that in order to estimate fish abundance, it is necessary to know the soniferous species of a given population, the daily and seasonal changes in individual call rates, and the relationships between acoustic parameters and fish abundance (ROUNTREE et al., 2006; GANNON, 2008). According to Desiderà et al. (2019), a careful analysis of the occurrence of different sound types and relative abundance is needed to understand the relationship between acoustic and taxonomic diversity.

Some studies have found a correlation between biodiversity metrics and acoustic diversity, by applying indices. For example, Desiderá et al. (2019) , although employing a short recording period, found a strong positive relationship between taxonomic diversity and acoustic diversity by applying indices of richness, diversity, and similarity (indices that were not ecoacoustic). Harris et al. (2016) found that two ecoacoustic indices (acoustic complex index—ACI, and acoustic entropy index—H) may be related to biodiversity in temperate reefs. Peck et al. (2021) found a correlation between abundance and species richness using ecoacoustic indices. Rice et al. (2017) tested three ecoacoustic indices (H, ACI, and acoustic diversity index—ADI) to compare the diurnal and nocturnal acoustic habits of fish. Their results showed a correspondence between ACI and the choruses and a negative correlation between H, ADI, and the choruses. However, they claim that indices are a promising tool for exploring diel or seasonal patterns of fish sounds. Despite being a promising tool, the use of acoustic indices in marine environments is at an early stage and needs further studies, as highlighted in recent reviews (MOONEY et al., 2020; PIERETTI; Danovaro, 2020; MINELLO et al., 2021). It is worth noting that these indices were developed for terrestrial environments and then directly applied to marine ecosystems (PIERETTI; DANOVARO, 2020). One challenge concerns their ability to discriminate the contribution of specific acoustic events in large datasets (RICE et al., 2017) and to distinguish between different biological sounds with similar patterns (MANCUSI et al., 2022).

Assessment of fish sound diversity is essential for mapping the spatial and temporal patterns of soniferous fish (DI IORIO et al., 2021). Understanding and monitoring the daily and seasonal variation of biological sounds and their relationship with the environment is fundamental for assessing the health and status of an ecosystem (RUPPÉ et al., 2015; CARRIÇO et al., 2020b; PIERETTI; DANOVARO, 2020). Long-term recordings also provide insights into important ecological information about species, such as acoustic social interactions and the effects of environmental parameters and anthropophony (VIEIRA et al., 2015).

4. MACHINE LEARNING

Due to the rapid development of passive acoustic techniques, recording systems are now capable of storing large amounts of data and with higher sampling frequencies, making manual analysis unfeasible (MOONEY et al., 2020; WADDELL et al., 2021; PARSONS et al., 2022). Therefore, data processing of large datasets requires robust computational methods (BIANCO et al., 2019; MUNGER et al., 2022). Artificial intelligence techniques are equally developing, and they offer a solution for automating tasks such as the detection, classification, and identification of biological acoustic signals.

Machine learning (ML), including deep learning (DL), is rapidly developing with significant results in different fields, including ecology (TUIA et al., 2022; PICHLER; HARTING, 2023). Applications in acoustics include source localization (NIU et al., 2017; LIU et al., 2020), sonar target recognition (KAMAL et al., 2013), geoacoustic inversion problems (SMARAGDAKIS; TAROUDAKIS, 2020), environmental sounds (WANG et al., 2006; BOUNTOURAKIS et al., 2015), human speech analysis (RICCARDI; HAKKANI-TUR, 2005; GAIKWAD et al., 2010), and music classification (HERRERA-BOYER et al., 2003; LEE et al., 2009). In the field of bioacoustics, these methods have been successfully applied in terrestrial environments for animal sound classification, such as insects (SILVA et al., 2013), frogs (HUANG et al., 2009), birds (BRAVO-SANCHEZ et al., 2021; MEHYADIN et al., 2021), bats (PARSONS; JONES, 2000), and other mammals (PANDEYA; LEE, 2018; CLINK; KLINCK, 2021). These techniques have also been used to analyse the vocalizations of farm livestock species, with the potential to monitor animal welfare (MCLOUGHLIN et al., 2019).

In the marine environment, artificial intelligence has been used to detect and classify cetacean sounds (ESFAHANIAN et al., 2014a, b; BARKLEY et al., 2019; ZHANG et al., 2019; SHIU et al., 2020; USMAN et al., 2020; FRASIER, 2021; ALLEN et al., 2021). Recently, several studies have emerged using machine learning and deep learning for fish sound detection and classification (RUIZ-BLAIZ et al., 2012; VIEIRA et al., 2015; KOTTEGE et al., 2015; NODA et al., 2016; SATTAR et al., 2016; DEMERTZIS et al., 2017; HARAKAWA et al., 2018; IBRAHIM et al., 2018a, b; LIN et al., 2018 ; MALFANTE et al., 2018a, b; KUMAR; JOSE, 2019; MONCZAK et al., 2019; VIEIRA et al., 2019; CHÉRUBIN et al., 2020; IBRAHIM et al., 2020; GUYOT et al., 2021; WADDELL et al., 2021; LAPLANTE et al., 2022; MANCUSI et al., 2022; MUNGER et al., 2022; MAHALE et al., 2023) (Table 1).

TABELA 1 - Publications using machine learning and deep learning techniques for fish sound detection, classification, and identification.

Task	Fish	Sound types	Technique	Features	Reference
Detection	*Cynoscion squamipinnis*, Sciaenidae	Stridulation trains	Unsupervised algorithms	Drumming frequency, STPL, LTPL, Timbral features[a]	RUIZ *et al*. (2012)
Detection and Identification	*Halobatrachus didactylus*, Batrachoididae	Boatwhistles, croaks, double croaks, grunts	HHM	b	VIEIRA *et al*. (2015)
Detection and Classification	*Tilapia mariae*, Cichlidae	Clicks	DA, LR	STFs, TTFs	KOTTEGE *et al*. (2015)
Identification	*Porichthys notatus*, Batrachoididae	Grunts, growls	SVM	MFCCs, LPCCs	SATTAR *et al*.(2016)
Identification	Unknown fish[c]		KNN, RF, SVM	MFCCs, LFCCs, SE, SL	NODA *et al*. (2016)

Identification	Unknown fish[c]		SCNNs	Features from spectral domain	DEMERTZIS *et al*. (2017)
Detection and Classification	Four grouper species, Epinephelidae[d]	Tonal calls, and pulse trains	KNN, SVM, Sparse classification	MRAFand MFCCs	IBRAHIM *et al*. (2018a)
Classification	Unknown fish[e]	Impulsions, drums, roars, and quacks	RF, SVM	84 features from time, frequency, and cepstral domains	MALFANTE *et al*. (2018a)
Detection and Classification	Unknown fish[e]	Impulsions, drums, roars, and quacks	CNN	84 features from time, frequency, and cepstral domains	MALFANTE *et al*. (2018b)
Classification	Four grouper species, Epinephelidae[d]	Tonal calls and pulse trains	CNN, LSTM		IBRAHIM *et al*. (2018b)
Detection	Unknown fish[e]	Pulses	PC-NMF, GMM	Features from spectral domain	LIN *et al*. (2018)
Detection and Classification	*Penahia argentata*, *Nibea mitsukurii*, Sciaenidae	Pulses	NLR, SVM, KNN	[f]	HARAKAWA *et al*. (2018)
Detection and Classification	Four Sciaenidae species[g]		[h]	SPL	MONCZAK *et al*. (2019)
Classification	[i]	Croaks, clicks, and grunts	DA, KNN, SVM	LWAP and AC	KUMAR & JOSE (2019)
Detection and Classification	*Argyrosomus regis*, Sciaenidae	Grunts and knocks	HMM	MFCCs, peak frequency, features from time domain	VIEIRA *et al*. (2019)
Detection	Three grouper species[j]	Tonal calls and pulse trains	CNN, LSTM	MFCCs	CHÉRUBIN *et al*. (2020)
Classification	Four grouper species, Epinephelidae[d]	Tonal calls and pulse trains	TL		IBRAHIM *et al*. (2020)
Detection	Species from genus *Alosa*, Clupeidae	'Bull' sounds[k]	CNN		GUYOT *et al*. (2021)
Detection and classification	Unknown fish[e]	Beats, buzz, croak, downsweep, pulse train, and jetski	NN	Features from time and frequency domains	WADDELL *et al*. (2021)
Identification	Pomacentridae	Pulse trains and chirps	CNN, TL	[l]	MUNGER *et al*. (2022)

Detection	Unknown fish[c]		DL		MANCUSI *et al*. (2022)
Detection and classification	Sciaena umbra		DL		LAPLANTE *et al.* (2022)
Classification	*Terapon theraps,* Sciaenidae species, and planctivorous fishes	Four fish sound types and an unknown 'buzz' sound	PCA, *K-means*	m	MAHALE *et al*. (2023)

STPL, short-term partial loudness; LTPL, long-term partial loudness; HMM, hidden Markov models; DA, discriminant analysis; LR, logistic regression; STFs, spectro-temporal features; TTFs, timbral texture features; SVM, support vector machine; MFCCs, mel frequency cepstral coefficients; LPCC, linear prediction cepstral coefficients; KNN, K-nearest neighbours; RF, random forest; LFCCs, linear frequency cepstral coefficients; SE, Shannon entropy; SL, syllable length; SCNN, spiking convolutional neural networks; MRAF, multi-resolution acoustic features; CNN, convolutional neural networks; LSTM, long short-term memory; PC-NMF, periodicity-coded non-negative matrix factorization; GMM, gaussian mixture models; NLR, novel logistic regression; LWAP, local wavelet acoustic pattern; AC, approximation coefficients; TL, transfer learning; NN, neural networks; DL, deep learning; PCA, principal component analysis.
a, Spectral centroid, cesptral coefficient, zero-crossing rate; b, energy, cepstrum, mel-frequency cepstral (MFC), perceptual linear predictive (PLP), delta coefficient, acceleration coefficient; c, fish sounds collected from internet databases; d, *Epinephelus guttatus*, *Epinephelus striatus*, *Mycteroperca venenosa*, *Mycteroperca bonaci*; e, dataset composed of sounds collected by the authors; f, frequency centroid, frequency band width, spectral rolloff, zero-crossing rate, RMS energy, MFCCs, polynomial features, auto correlation, RMS of amplitude, statistics of peaks, statistics of amplitude; g, *Pogonias cromis*, *Bairdiella chrysoura*, *Cynoscion nebulosus*, *Sciaenops ocellatus*; h, feature-based signal detector designed by the authors; i, sounds of three fish species collected from internet sound bases: *Micropogonias undulatus* (Sciaenidae), *Hippocampus erectus* (Syngnathidae), *Epinephelus adscensionis* (Epinephelidae); j, *Epinephelus guttatus*, *Epinephelus striatus*, *Mycteroperca venenosa*; k, sounds made with caudal fin during spawning; l, frequency bandwidth, spectral peak frequency, pulse repetition frequency, pulse number, inter-pulse-interval, call duration; m, twelve parameters, including features from time domain (call duration, number of pulses per call, inter-call durations), features from spectral domain (peak frequency and dominant peak PSDs), and other acoustic metrics (Acoustic Evenness Index—AEI, ACI, and SPLrms).

The fundamentals of ML consist of learning to recognize patterns and identify relationships between measured data (features), commonly using statistics, and generating an output of interest (target). A crucial step in machine learning is to divide the dataset into a training set and a test set. In general, the larger the training dataset with a greater number of sound examples, the higher the recognition accuracy of the model (VIEIRA et al., 2015). The two main groups of machine learning methods are unsupervised and supervised learning, with the latter category being the most widely used (BIANCO et al., 2019).

Supervised learning can solve regression or classification problems. Models are trained using labelled data, where observations are separated into classes of interest to find specific relations or structures in the input data, and correct outputs are generated by the trained data (MALFANTE et al., 2018a; BIANCO et al., 2019).

Although the labelling step is sometimes performed manually by experts, training algorithms with labelled data helps to reduce classification errors and improve the performance of the classification model (MALFANTE et al., 2018a). Common supervised algorithms include linear regression, logistic regression, naive Bayes, support vector machine (SVM), K-nearest neighbour (KNN), and random forest (RF) (CARUANA; NICULESCU-MIZIL, 2006; GOODFELLOW et al., 2016; BIANCO et al., 2019).

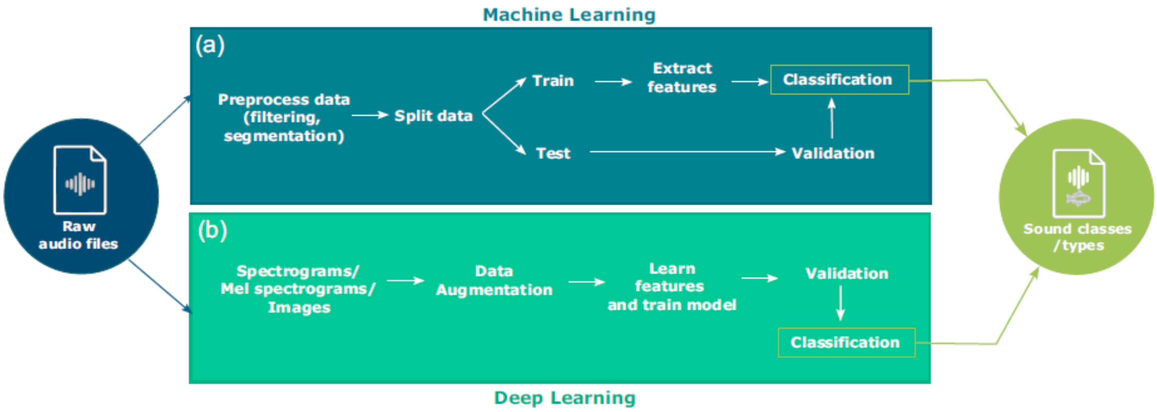
Supervised algorithms have been used to detect and classify fish sounds. For instance, Noda et al. (2016) have used KNN, RF, and SVM to identify 102 fish species with an average accuracy of 95.24, 93.56, and 95.58%, respectively. Malfante et al. (2018a) used RF and SVM on a dataset of 913 examples to identify four different fish sound types, achieving a classification performance of 96.9%. Harakawa et al. (2018) presented a hybrid approach that performs multistage classification. In this approach, novel logistic regression (NRL) performed adaptive feature weighting and provided complementary results, focusing on the accuracy of SVM and KNN classifiers.

Unsupervised learning consists of training models using unlabelled datasets to cluster or associate data samples based on their similarities and without supervision. The main goal of this approach is to learn the natural structure of a dataset. Unsupervised algorithms have been used to reduce the dataset dimensionality while keeping its structure (ROCH et al., 2021). Common unsupervised algorithms are Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), K-means, hierarchical clustering analysis, and principal component analysis (PCA) (BIANCO et al., 2019; ROCH et al., 2021).

Clustering approaches have been applied to identify different biological sources without labels (LIN et al., 2017). These algorithms are also employed in long-term recordings to investigate the variability in the soundscape, by identifying temporal and spectral patterns in the dataset. However, their performance can be affected by overlapping sounds or due to the low signal-to-noise ratio (SNR) recordings (LIN et al., 2018; LIN; TSAO, 2020). Ozanich et al. (2021) compared GMMs, conventional clustering, and deep embedded clustering for the task of discriminating fish calls from whale songs. In their approach, deep embedded clustering was able to distinguish simulated fish calls and whale songs, even when these signals overlapped. This method combines deep learning and clustering, and its performance is superior to conventional clustering (HUANG et al., 2023). MAHALE et al. (2023) used an unsupervised approach employing PCA (for dimensionality reduction) and K-means (to detect patterns in the dataset). The overall classification performance of four fish sounds was 89.84%. Malfante et al. (2018a) suggest that unsupervised learning is a means to inspect the variability within a class of sounds, as a class may have subclasses that could be related species or individuals. Lin et al. (2018) suggest a multi-method approach to address the variability in the datasets, applying supervised algorithms as a first step to detect fish calls, and then unsupervised algorithms to identify potential errors.

The recognition of biological sounds through machine learning techniques involves some basic steps (Figure 1). Raw acoustic data must first be pre-processed by filtering the dataset to select the frequencies of interest. Low-frequency bands are often used to detect fish sounds. For instance, Malfante et al. (2018a) used frequency-filtered signals with bandwidths of 50–450 Hz and 500–900 Hz. To compute biodiversity and classify fish sounds, Mahale et al. (2023) categorized three frequency bands: “full band”(50–22050 Hz), “low-frequency fish band”(100–2000 Hz), and “high-frequency shrimp band”(2000–20000 Hz). Pre-processing is also important for establishing detection thresholds (MALFANTE et al., 2018a). Methods such as matched filter and spectrogram correlation or energy thresholds can be applied to detect and extract fish calls in the dataset (ROWELL et al., 2017). The next step is feature extraction, to extract temporal and spectral patterns of the detected sounds (MALFANTE et al., 2018a; WADDELL et al., 2021). Some studies also used a segmentation phase before feature extraction, with the automatic splitting of signals into syllables (NODA et al., 2016; MAHALE et al., 2023).

FIGURA 1 - Workflow for a generic fish sound classifier using machine learning (a) and deep learning (b). The main characteristic of deep learning is automatic feature extraction. More commonly in deep learning approaches, a data augmentation step can be included.



The performance of the ML models is crucial for detection and classification problems. Some nontrivial problems may occur, such as bias (when a learner is unable to learn distributions or separate boundaries) and variance (error due to a specific training set) (ROCH et al., 2021). There are also other challenges in machine learning, such as underfitting and overfitting. Underfitting occurs when the model is unable to find relationships between variables in the training set, leading to a large error. Overfitting means that the model has a small error in the training set and a large error in the test set, resulting in a decrease in the generalization properties of the model (e.g. the ability to perform well in the presence of novel inputs) (GOODFELLOW et al., 2016; JABBAR; KHAN, 2015). Therefore, ML algorithms require evaluation and validation tasks, using metrics such as accuracy, precision, and recall. Cross-validation is used to evaluate the performance of the model by checking its accuracy multiple times, which can lead to a higher reliability of the results (RAMEZAN et al., 2019). Most of the studies presented here had an average detection or classification performance between ~80 and 95%. Accuracy can be influenced by factors such as the technique adopted, the size of the training set, datasets with low SNR, and fish call types (USMAN et al., 2020; WADDELL et al., 2021).

The use of machine learning for fish sound classification is very recent and faces some limitations (MALFANTE et al., 2018a). Currently, machine learning has contributed significantly to the automation of processes and improving the detection, classification, and identification of fish calls, even in noisy recordings. However, this has only been done for a few target species within a dataset, and most species do not have their repertoire documented (PARSONS et al., 2022). To effectively apply automatic fish species identification by sound, it is necessary to characterize the vocal repertoire of the species. One approach is to perform unsupervised learning for call repertoire analysis. By applying clustering algorithms, the dataset can be grouped into clusters with similar temporal and spectral characteristics (LIN et al., 2018).

Nevertheless, machine learning has shown advantages over the manual analysis of sounds, with promising results and applications in fish sound identification, even under varying acoustic conditions (MUNGER et al., 2022). Some of the systems developed can be applied to larger datasets of continuous recordings, performing real-time analysis of sounds (MALFANTE et al., 2018a), and can be adapted for multiclass identification with different call types (SATTAR et al., 2016). ML has also demonstrated the potential to detect and validate novel sounds from the same species, thus contributing to the description of vocal repertoires. For instance, Vieira et al. (2019) have shown that *Argyrosomus regius* has a wider repertoire than previously known, with variability of “grunts” and “knocks”. In addition, artificial intelligence techniques are rapidly developing, and deep learning has shown potential in a variety of fields (STOWELL, 2022), making it a promising approach for fish sound identification (IBRAHIM et al., 2020; MUNGER et al., 2022).

4.1. Feature selection

Because audio signals have high dimensionality, there is a need for reduction and feature extraction, which is a key step in sound analysis (ABAYOMI-ALLI et al., 2022). Different acoustic features can be selected from spectrograms to assess the acoustic activity and behaviour of soniferous fish. The main goal of this step is to obtain temporal and spectral patterns to characterize the target sounds (NODA et al., 2016). These characteristics contain information that allows the discrimination between the sound types and species (MOONEY et al., 2020). Thus, classical approaches rely on the analysis of a time series of temporal and spectral patterns of spectrograms to infer which species are present in the dataset.

In machine learning approaches, any algorithm requires the transformation of the data into features (ROCH et al., 2021). The way in which acoustic signals are represented is fundamental to sound analysis using ML (MALFANTE et al., 2018a). Feature extraction is one of the most important steps in the ML process since the algorithms depend on the features for training and test data (SHARMA et al., 2020). Currently, there are several methods for feature extraction, which perform various transformations on the data to extract relevant measures. The most popular transformations are the Discrete Fourier Transform (DFT) and the Fast Fourier Transform (FFT). These algorithms are applied to transform a signal from the time domain to the time-frequency domain or frequency domain, respectively (MCLOUGHLIN et al., 2019).

In order to detect and classify different types of biological sounds, researchers often employ multiple acoustic features as inputs to the learning algorithms. Common features include metrics from the time, frequency, and cepstral domains. Features from the cepstral domain consist of obtaining cepstral coefficients from sound (MALFANTE et al., 2018a; KUMAR; JOSE, 2019). Mel frequency cepstral coefficients (MFFCs) are widely used in sound classification, especially in the field of speech recognition (MCLOUGHLIN et al., 2019). This feature is based on the sensitivity of human perception to frequency differences since it is derived from a psychoacoustic scale, the mel scale (SARANGI; SAHA, 2012).

Besides MFCCs, other features were found in recent work, such as linear prediction cepstral coefficients (LPCCs), short-term partial loudness (STPL), long-term partial loudness (LTPL), timbral features, energy, cepstrum, spectro-temporal features (STFs), and timbral texture features (TTFs), among many others (Table 1). For example, Ruiz-Blais et al. (2012) obtained timbre measures by using features such as cepstrum, spectral centroid, and zero crossings to detect sciaenid calls. Noda et al. (2016) used MFCCs to identify fish species. The authors state that MFCCs outperformed LFCCs because fish produce sounds mainly at lower frequencies, and MFCCs have a higher resolution in these frequency bands. Malfante et al. (2018) used 84 features (from time, frequency, and cepstral domains) and found four different classes of fish sounds. This large feature set was used to capture different signal properties and was inspired by works from different fields, such as speech recognition, music genre and instrument classification, and sonar sound classification.

The process of feature selection can be considered a research task itself (MCLOUGHLIN et al., 2019). There is a myriad of acoustic features to be explored; however, most of them were designed for different audio domains. Thus, there is a need to select appropriate features for fish acoustic signals, taking into account their specificities. The process of choosing which features are more adequate for modelling these sounds can be enhanced by using feature importance, which ranks features according to their contribution of each one to the model. Some supervised algorithms, such as random forest and extra trees can estimate feature importance (AHMAD et al., 2018). Besides, the feature selection and the classification method can determine the accuracy of the systems (JASIM et al., 2021).

5. DEEP LEARNING

Deep learning (DL) contributes to the automation of tasks in different fields. In the acoustic field, it contributes to sound event detection and source localization (BIANCO et al., 2019). DL has promoted significant progress in audio event detection (GUYOT et al., 2021). The field of bioacoustics also benefits from DL, as it helps to process large datasets from field recordings (PARSONS et al., 2022; STOWELL, 2022). The DL algorithms used in bioacoustics are derived from those used in the audio domain, such as speech and music analysis; however, bioacoustics has its own specificities and requirements (STOWELL, 2022). Nevertheless, these methods have proved to be effective for the detection and classification of animal vocalizations with high precision and recall (SHIU et al., 2020). Furthermore, DL has outperformed manual approaches to blue whale calls, even with experienced analysts, with higher recall and fewer false positives (MILLER et al., 2023). Fish sound detection and classification studies in the last two years have used mostly DL rather than ML approaches (CHÉRUBIN et al., 2020; GUYOT et al., 2021; LAPLANTE et al., 2022; MANCUSI et al., 2022 ; MUNGER et al., 2022).

DL consists of a set of supervised and unsupervised ML algorithms that can perform tasks from large datasets (GOODFELLOW et al., 2016). It is considered to be synonymous with deep neural networks (DNNs), as DL concerns neural networks

(NNs) with multiple layers, which include input, output, and more than one hidden layer (GOODWIN et al., 2022). A key characteristic of DL approaches is the automatic feature extraction (Figure 1). Whereas in ML the feature extraction is performed manually, DL can reduce the effort for this task since the models learn hierarchical features from raw datasets (GOODFELLOW et al., 2016; GOODWIN et al., 2022). Another notable contribution of DL, which makes it popular in sound analysis, is its good performance under noisy conditions (MOONEY et al., 2020). DL can also reduce computational efforts required for sound classification using supervised approaches, such as SVM. Recent developments in DL can overcome these limitations by learning a non-linear mapping of the inputs from the data (BIANCO et al., 2019). These advances allow DL algorithms to use complex non-linear functions to learn features at different levels in both supervised and unsupervised approaches (GOODWIN et al., 2022). DNNs have also reduced classification errors and overfitting in NNs, and they are being used to detect patterns in the bioacoustics field (BIANCO et al., 2019).

As well as in ML, DL presents different architectures. One popular architecture for supervised learning is convolutional neural networks (CNNs). This architecture was first introduced for image processing, and it has achieved the state-of-the-art tasks such as image and speech recognition, semantic image segmentation, and natural language processing, among others (AHMAD et al., 2019). It consists of self-learning significant features from the training dataset through convolution layers by using filters and tuning their coefficients during the training step. It learns basic features in the lower layers, and the more complex ones as the data propagate to the deeper layers (AHMAD et al., 2019). The advantage of this architecture is the reduction in data pre-processing and the parameters of the model, which diminishes computational burden (BIANCO et al., 2019; SALEH et al., 2022).

CNNs have recently been used for underwater sound classification. For example, Mishachandar and Vairamuthu (2021) applied them to classify ocean sounds within a wide frequency range. They found six classes of sounds: anthropogenic and natural sounds; vocalizations of cetaceans, fish, and invertebrates; and unidentified sounds. CNNs have also been used to detect and classify cetacean sounds (BERMANT et al., 2019; JIANG et al., 2019; SHIU et al., 2020; ALLEN et al., 2021). A common approach to audio classification is to convert spectrograms into a visual representation and use image classification, as CNNs are effective at learning patterns from images (IBRAHIM et al., 2018b; MISHACHANDAR; VAIRAMUTHU, 2021). In the computer vision domain, the models can directly extract features from images (ZHAO et al., 2018). Recent studies on fish sound detection and classification have used CNNs (DEMERTZIS et al., 2017; IBRAHIM et al., 2018b; MALFANTE et al., 2018b; CHÉRUBIN et al., 2020; MUNGER et al., 2022). Most of them used image classification in the spectrograms. For instance, Munger et al. (2022) converted sound detection and classification tasks into a supervised approach for image recognition.

One limitation of CNNs is the need for a large labelled dataset to train the networks (ZHUANG et al., 2020). As in many situations, labelled data are not available; an alternative is transfer learning (TL). This is another popular DL architecture, consisting of DNNs previously trained, that transfers the knowledge learnt to a new dataset (AHMAD et al., 2019; ZHUANG et al., 2020). It uses the previously trained weights of a model to initiate the parameters of new models, adjusting them to the new domain (AHMAD et al., 2019). This makes model training processes faster, with the benefit of improving their performance and requiring smaller datasets (MUNGER et al., 2022). With the goal of classifying grouper species and vessel sounds, Ibrahim et al. (2020) applied TL to extract fish sound features from spectrograms and scalograms (time–frequency representations of a signal by wavelet transform), by using DNNs as feature extractors. Their results show that TL outperformed the manually identified features approach, demonstrating the effectiveness of this approach.

6. CHALLENGES AND LIMITATIONS

Machine learning and deep learning techniques are rapidly evolving, making it possible to process the increasing amount of PAM data. While these techniques are comparatively advanced in the field of speech recognition, they are still in their infancy in the field of fish sound recognition (MALFANTE et al., 2018a). Currently, none of the acoustic methods can be considered as an ideal tool for taxonomic fish recognition, with applications to biodiversity estimation (NODA et al., 2016;

MANCUSI et al., 2022). However, other methods also have disadvantages and limitations, as discussed in the previous sections of this manuscript.

The remarkable diversity of sounds produced by fish and the complexity of the underwater acoustic environment pose a challenge to automatic detection and classification systems. However, the main barrier to fish sound recognition through artificial intelligence is clearly the lack of species-specific validated sound data and the fact that very few species have their vocal repertoires fully characterized. ML and DL have shown potential in the acoustics field; however, it is necessary a previous knowledge of the sound sources and data for training (MOONEY et al., STOWELL et al., 2019a), with different examples for the same species. Most of the studies presented here succeed in detecting and classifying fish sounds with high accuracy. Nevertheless, the attribution of these sounds to specific species is still challenging, due to the lack of data to compare the detected sounds.

In addition to the lack of validated sounds, the intraspecific variability of fish acoustic signals also complicates the classification and identification tasks. Variability in spectral and temporal patterns of the same species and individuals can occur as a function of size (MYRBERG et al., 1993; LOBEL; MANN, 1995; CONNAUGHTON et al., 2000), ontogeny, and sexual dimorphism (AMORIM, 2006; KASUMYAN, 2009). Variation due to geographic localization (FINE, 1978; MANN; LOBEL, 1998; PARMENTIER et al., 2005; PHILIPS; JOHNSTON, 2008), seasonality, and reproductive contexts are also reported (AMORIM, 2006; VERZIJDEN et al., 2010; VIEIRA et al., 2019).

Other factors that may influence fish sound detection are related to changes in abiotic factors, such as temperature (LADICH, 2018), which affect fish sound production. For instance, Sattar et al. (2016) demonstrated a relationship between ocean temperature and classification accuracy of *Porichthys notatus*, as the call frequency of this species is correlated with temperature. Vieira et al. (2019) reported seasonal and diel changes in sound production, and variations in acoustic features (pulse period, fundamental frequency, and peak frequency), due to changes in temperature. Monczak et al. (2019) found that the call rates of black drum, silver seatrout, and spotted seatrout increased with increasing temperature while call rates of red drum increased with decreasing temperature. Chérubin et al . (2020) found no significant correlation between calling rate patterns and environmental parameters.

However, temperature and salinity were uniform during their study, which was conducted in a well-mixed environment. Munger et al. (2022) found that an increase in tidal amplitude was associated with an increase in the presence of damselfish sounds, as the lunar cycle influences both tides and the reproductive activities of this group of fish.

Understanding the propagation conditions is crucial for efficient fish sound detection, as sound is attenuated as it travels from the source (fish) to the receiver (hydrophones). Some authors have considered the influence of propagation conditions in automatic detection and classification systems. Harakawa et al. (2018) have used multiple classifiers adapted according to the underwater environment, as the performance of features and classifiers depends on the environmental conditions. Chérubin et al. (2020) estimated a detection threshold using the passive sonar equation. The transmission loss (TL) parameter of the equation was estimated using sound velocity profiles. Munger et al. (2022) used conservative propagation estimates to calculate the detection range. In this context, the propagation conditions can be useful parameters for automatic detection and classification systems. Acoustic propagation models can be applied to simulate complex physical processes as sound travels through the underwater environment. An output of propagation models is the TL of a particular frequency as a function of depth and distance from source to receiver (ETTER, 2003). This is fundamental for understanding which frequencies of sound emitted by fish would be attenuated in a particular area, hence estimating more accurate detection ranges. The use of specific training datasets that reflect the unique local acoustic conditions are also encouraged (MUNGER et al., 2022).

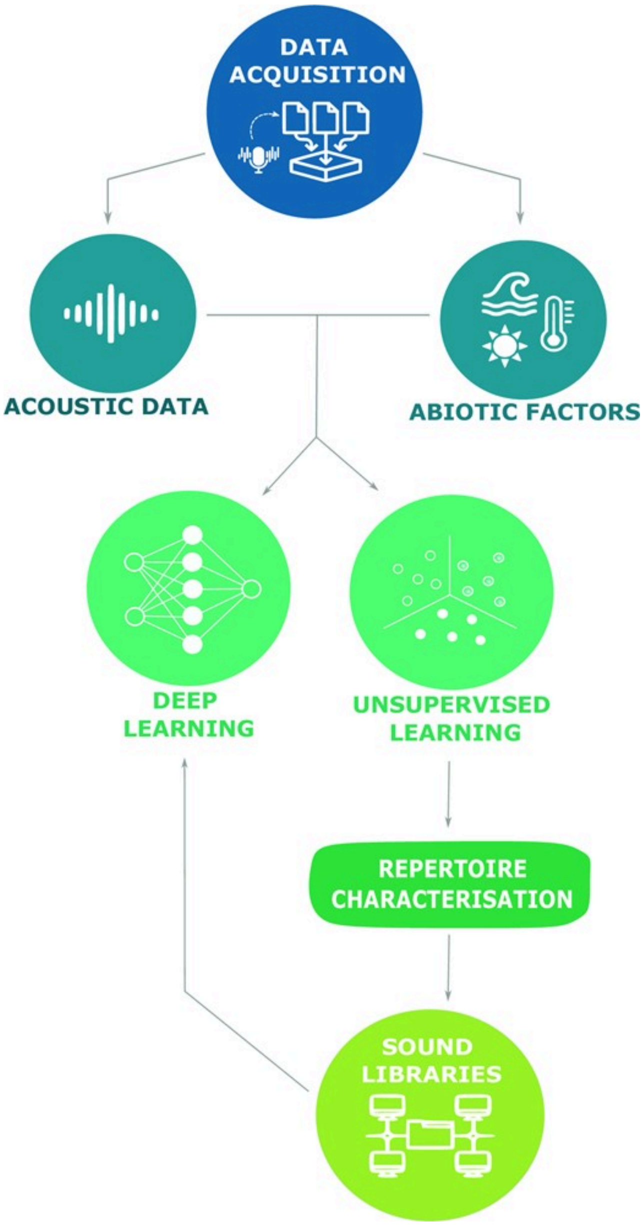
The complexity of underwater soundscapes can also affect the robustness and generalization of ML and DL algorithms due to overlap and masking effects from other biological sources, as well as geophonic and anthropophonic sources (STOWELL et al., 2019b; MUNGER et al., 2022). Some biotic sources that produce choruses often dominate soundscapes, which complicates separating target sounds. The dominant frequencies of sounds from anthropogenic sources often produce spectral overlapping in biological sounds (CROVO et al., 2022). Many fish species produce sounds in a frequency range that comprises most anthropogenic and geophonic sources (HAWKINS; POPPER, 2018). Instead of masking, increases in

SPL may occur in the presence of geophonic sources (MUNGER et al., 2022), and anthropogenic sources (e.g. Lombard effect, including changes in fish call patterns) (HOLT; JOHNSTON, 2014). Fish sound production may also be inhibited in the presence of predators (LADICH, 2022), which can influence the detection problem. Feature extraction techniques can also be affected by noise, reducing the classification accuracy (USMAN et al., 2020). As soundscapes are highly dynamic and fish sound production follows diel and seasonal patterns, long-term recordings are recommended to provide datasets that capture the variability of target signals (MALFANTE et al., 2018a; VIEIRA et al., 2019; STOWELL et al., 2019a; MOONEY et al., 2020).

7. FUTURE STEPS

Fish sound recognition via passive acoustics has been improved by recent artificial intelligence techniques. Some factors can influence the performance of ML and DL algorithms. The choice of the technique will depend on the characteristics of fish acoustic signals and the available dataset. The robustness of the algorithms can be increased by selecting an appropriate feature set. The increasingly strong performance of automatic detection and classification in the bioacoustics field has benefited in part from large labelled datasets for training (STOWELL et al., 2019a). However, although many fish are soniferous, most sounds have not been associated to these species (MOUY et al., 2018). Furthermore, there is a lack of annotated data for training supervised methods and CNNs. One approach to address the lack of validated data is to use unsupervised techniques to separate different sound sources (MOONEY et al., 2020). This can help identify groups of sounds that are similar, thus helping to identify new species based on common acoustic patterns of known species. Here we highlight some points that would help design future studies (Figure 2).

FIGURA 2 - Overview of the application of AI to fish sound detection, classification, and identification, with key topics for future research.



7.1. Data augmentation

The species-specific calls and the variability of intraspecific sounds are crucial for the recognition of individual species (VIEIRA et al., 2019). However, the lack of species-specific validated calls and repertoires fully characterized for most species is a limitation for the identification of individual species in a multi-species dataset. Modifications to data samples, such as frequency shifting and warping in the frequency and time domains, may change subtle characteristics such as those that allow species discrimination (STOWELL, 2022). Data augmentation is a technique that can artificially increase the size of training datasets by performing slight modifications in the sampled data and creating new examples (STOWELL et al., 2019a).

This technique is an approach to dealing with the scarcity of labelled data for training machine learning and deep learning algorithms. Data augmentation helps with generalization, and makes the dataset more diverse (SHIU et al., 2020; STOWELL, 2022). In the audio domain, the modifications can include frequency and time shifts, warping, and the addition of low-amplitude noise, which can be directly applied to the raw data or in the spectrograms (LASSECK, 2018; NANNI et al., 2020; ABAYOMI-ALLI et al., 2022). Data augmentation can also be implemented in the image representations of the audio by rotation, translation, scaling, or reflection (NANNI et al., 2020).

Besides improving the performance in training sound detection and classification models for small datasets, data augmentation can enhance their robustness to the variation in target signals and the presence of noise and overlapping vocalizations (KAHL et al., 2021; SHIU et al., 2021; ABAYOMI-ALLI et al., 2022). Moreover, data augmentation applied to a specific problem domain can be helpful in dealing with unexpected variations in data and enhance generalization to new recording conditions and habitats (LASSECK, 2018; KAHL et al., 2021). Data augmentation also helps with the data imbalance, which affects the performance of deep learning algorithms (ABAYOMI-ALLI et al., 2022). As unbalanced data are common in the bioacoustics field (STOWELL, 2022), the use of data augmentation has recently increased for the detection and classification of animal sounds in terrestrial (LASSECK, 2018; STOWELL et al., 2019a; NANNI et al., 2020; KAHL et al., 2021) and in marine environments (PADOVESE et al., 2021; SHIU et al., 2021; CAI et al., 2022). Among the reviewed papers, two studies have used data augmentation. For the classification of grouper species, Ibrahim et al. (2020) increased the number of samples of a particular species by applying rotation and scaling to the converted images of spectrograms. Waddell et al. (2021) achieved a higher accuracy by adjusting the scaling and translation of the images of different fish call types.

7.2. Sound libraries

The lack of labelled data to train algorithms is a limitation to supervised approaches and CNNs (PARSONS et al., 2022), as there is a great amount of unknown sounds that have not been matched to a particular species (MOUY et al., 2018). CNNs are capable of learning features, but they require previously trained models. Even unsupervised approaches, such as HMMs, can be trained in a supervised manner. Therefore, it is often necessary to label, train, or compare the data (TAMPOSIS et al., 2019). Preparing large datasets for training is a difficult and time-consuming task (HARAKAWA et al., 2018), and finding these datasets is equally challenging. However, the main limitation is the lack of data to compare and validate the calls detected and classified by automated systems.

While there are many datasets for training in the human speech and music domains, there is a lack of a systematic fish sound database with different examples of fish calls for the same species (MANCUSI et al., 2022). Thus, there is a need for reliable and comprehensive databases (MOONEY et al., 2020; USMAN et al., 2020; PARSONS et al., 2022). Sound libraries can be an important tool for comparing sounds within and between species, and for confirming whether a classified sound has already been identified. Some fish sound databases are available, such as the GSO Fish Sounds Collection, with examples of fish sounds collected by Fish and Mowbray (1970), and Fish Sounds (LOOBY et al., 2022b), and The Sound Table (of Fish Base) (FROESE; PAULY, 2022), which are complementary projects collaborating since 2021. However, most species are not catalogued and various sound types are not described in the existing libraries.

Recently, Parsons et al. (2022) have highlighted the urgent need for a global sound repository that integrates and expands existing libraries. This comprehensive database would provide labelled data for training ML and DL algorithms, as well as a large amount of unanalysed data to help identify unknown sounds from different species rather than the target species. Currently, it is not possible to know how many fish species are vocal, on a global scale (PARMENTIER et al., 2021). A database containing unidentified sounds would help to identify and increase the number of known vocal fish species. The first step is to generate these validated data to feed this database, and an alternative could be applying integrative approaches.

Combining video and acoustic recordings can help to develop a catalogue of fish sounds (MOUY et al., 2018, 2023).

7.3. Source separation

Although many studies focus on a few species of interest, acoustic recordings contain several sounds from the local soundscape (PARSONS et al., 2022), including many other species and non-biological sounds. The presence of multiple and simultaneous sources hampers the extraction of information about particular sounds. Thus, to analyse the target species sounds, it is necessary to reduce noise and separate these sounds from the soundscape. Traditional approaches to source separation in single-channel audio are generally based on matrix decomposition methods (TOLKOVA; KLINCK, 2022). Filtering is typically the first step in separating the frequency range of interest. However, despite knowing the representative frequency bands of a particular species, other sounds will surely interfere (LIN; TSAO, 2020). Wavelets can be applied to the denoising problem, improving SNR in environments with a wide range of noise (IBRAHIM et al., 2018b). Another common approach is applying statistics-based techniques for noise reduction. Nonetheless, these techniques are not appropriate for non-stationary noise (LIN; TSAO, 2020). Some authors recommend blind source separation (BSS), another class of denoising algorithms, widely used in speech and music applications (Lin et al., 2018; LIN; TSAO, 2020).

Sound detection and classification systems must be able to distinguish target sounds from background noise, hence providing accurate classification (PARSONS et al., 2022). Source separation also facilitates the feature extraction task (LIN; TSAO, 2020). The constraints in the identification of fish sounds under noisy conditions are recurrent. However, combining supervised and unsupervised approaches can optimize these tasks (LIN et al., 2018). Lin et al. (2017) proposed an unsupervised separation of biological sounds in long-term recordings of terrestrial and marine datasets using a periodicity-coded non-negative matrix factorization (PC-NMF). This approach enabled noise reduction and highlighted biological choruses without training datasets in a supervised manner.

A training approach employed in speech recognition is to use clean datasets and then add noise, so that the models can learn the statistical difference between clean and noisy data (LU et al., 2013). Mancusi et al. (2022) trained supervised networks for separating fish sounds from background noise. In their approach, the authors created a dataset for source separation by overlapping fish calls with recorded ambient noise from different sites. Another approach for separating overlapping biological sounds (e.g. bird calls in terrestrial environments) was proposed by Tolkova and Klinck (2022), using an acoustic vector sensor (AVS). More recently, a toolbox for supervised and unsupervised source separation based on NMF has been developed by Sun et al. (2022), aiming to help explore the acoustic diversity within the soundscapes. Besides reducing noise and improving the detection of sounds of interest, the toolbox allows the identification of biotic and abiotic sounds and helps to investigate the repertoire of target species. Kim et al. (2023) developed an automated tool for separating fish choruses from underwater soundscapes. Called SoundScape Learning (SSL), it consists of an integrated unsupervised method employing randomized robust principal component analysis (RRPCA), clustering, and an NN. Although the identification of the chorus to species level was not possible, the method is robust to long time scales and different soundscapes, with potential for application to other marine and terrestrial datasets.

8. CONCLUSIONS

This paper reviews the use of artificial intelligence for fish sound identification and characterization. Machine learning, including deep learning, is producing increasingly accurate results for fish sound recognition. ML and DL algorithms have been fundamental in processing long-term datasets and automating the detection, classification, and identification of fish sounds. The choice of the technique must consider the nature of fish acoustic signals and the available datasets. Selecting an appropriate feature set can improve the performance of the algorithms. The main challenge in applying AI to recognize fish sounds is the lack of validated sound data for individual species and the fact that only a small number of species have their vocal repertoires thoroughly characterized. Other factors that may affect the robustness of the algorithms include the abiotic factors that influence both fish acoustic behaviour and sound propagation, and the characteristics and variability of the soundscapes. The choice of a particular technique must therefore take into account the specificities of each problem and reflect local conditions. In this review, we have highlighted some points to be explored in future research, including source separation as well as data augmentation and the need for a systematic fish sound library to address the lack of species-specific validated data of fish sounds. Once filling these knowledge gaps, the combination of passive acoustics and artificial intelligence can help to make PAM an effective monitoring tool with applications in the management and conservation of fish communities.

9. REFERENCES

ABAYOMI-ALLI, O. O.; DAMAŠEVICIUS, R.; QAZI, A.; ADEDOYIN-OLOWE, M.; MISRA, S. Data augmentation and deep learning methods in sound classification: a systematic review. *Electronics*, v. 11, p. 3795, 2022.

ACKERMAN, J. L.; BELLWOOD, D. R. Comparative efficiency of clove oil and rotenone for sampling tropical reef fish assemblages. *Journal of Fish Biology*, v. 60, p. 893–901, 2002.

AHMAD, M. W.; REYNOLDS, J.; REZGUI, Y. Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production*, v. 203, p. 810–821, 2018.

AHMAD, J.; FARMAN, H.; JAN, Z. Deep learning methods and applications. In: KHAN, M.; JAN, B.; FARMAN, H. (ed.). *Deep learning: convergence to big data analytics*. Springer, The Gateway, p. 31–42.

ALLEN, A. N.; HARVEY, M.; HARRELL, L.; JANSEN, A.; MERKENS, K. P.; WALL, C. C.; CATTIAU, J. et al. A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Frontiers in Marine Science*, v. 8, p. 607321, 2021.

AMORIM, M. C. P. Diversity of sound production in fish. In: LADICH, F.; COLLIN, S. P.; MOYA, M. L.; KAPOOR, B. G. (ed.). *Communication in fishes*. v. 1, p. 71–104, 2006.

BARKLEY, Y., OLESON, E.M., OSWALD, J.N., and FRANKLIN, E.C. 2019. Whistle classification of sympatric false killer whale populations in Hawaiian waters yields low accuracy rates. *Frontiers in Marine Science*, 6: 645.

BERMANT, P.C., BRONSTEIN, M.M., WOOD, R.J., GERO, S., and GRUBER, D.F. 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific Reports*, 9: 1–10.

BIANCO, M.J., GERSTOFT, P., TRAER, J., OZANICH, E., ROCH, M.A., GANNOT, S., and DELEDALLE, C.A. 2019. Machine learning in acoustics: theory and applications. The Journal of the Acoustical Society of America, 146: 3590–3628.

BOUNTOURAKIS, V.; VRYSIS, L.; PAPANIKOLAOU, G. Machine learning algorithms for environmental sound recognition: towards soundscape semantics. In: AUDIO MOSTLY 2015 – INTERACTION WITH SOUND, 2015, Thessaloniki. *Proceedings…* Thessaloniki: ACM, 2015. p. 1–7.

BRAVO SANCHEZ, F. J.; HOSSAIN, M. R.; ENGLISH, N. B.; MOORE, S. T. Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. *Scientific Reports*, v. 11, p. 1–12, 2021.

BURCHARDT, L. S.; PICCIULIN, M.; PARMENTIER, E.; BOLGAN, M. A primer on rhythm quantification for fish sounds: a Mediterranean case study. *Royal Society Open Science*, v. 8, p. 210494, 2021.

CAI, W.; ZHU, J.; ZHANG, M.; YANG, Y. A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation. *Sensors*, v. 22, p. 7443, 2022.

CARRIÇO, R.; SILVA, M. A.; MENEZES, G. M.; VIEIRA, M.; BOLGAN, M.; FONSECA, P. J.; AMORIM, M. C. P. Temporal dynamics in diversity patterns of fish sound production in the Condor seamount (Azores, NE Atlantic). *Deep Sea Research Part I: Oceanographic Research Papers*, v. 164, p. 103357, 2020a.

CARRIÇO, R.; SILVA, M. A.; VIEIRA, M.; AFONSO, P.; MENEZES, G. M.; FONSECA, P. J.; AMORIM, M. C. P. The use of soundscapes to monitor fish communities: meaningful graphical representations differ with acoustic environment. *Acoustics*, v. 2, p. 382–398, 2020b.

CARUANA, R.; NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 23., 2006, Pittsburgh, PA. *Proceedings…* Pittsburgh: ACM, 2006. p. 161–168.

CHEAL, A. J.; EMSLIE, M. J.; CURREY-RANDALL, L. M.; HEUPEL, M. R. Comparability and complementarity of reef fish measures from underwater visual census (UVC) and baited remote underwater video stations (BRUVS). *Journal of Environmental Management*, v. 289, p. 112375, 2021.

CHÉRUBIN, L. M.; DALGLEISH, F.; IBRAHIM, A. K.; SCHÄRER-UMPIERRE, M.; NEMETH, R. S.; MATTHEWS, A.; APPELDOORN, R. Fish spawning aggregations dynamics as inferred from a novel, persistent presence robotic approach. *Frontiers in Marine Science*, v. 6, p. 779, 2020.

CLINK, D. J.; KLINCK, H. Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. *Methods in Ecology and Evolution*, v. 12, p. 328–341, 2021.

CONNAUGHTON, M. A.; TAYLOR, M. H.; FINE, M. L. Effects of fish size and temperature on weakfish disturbance calls: implications for the mechanism of sound generation. *Journal of Experimental Biology*, v. 203, p. 1503–1512, 2000.

CROVO, J. A.; MENDONÇA, M. T.; JOHNSTON, C. E. Acoustic modulation of reproductive hormones in the blacktail shiner, *Cyprinella venusta*, a soniferous cyprinid. *Animal Behaviour*, v. 186, p. 101–106, 2022.

DEMERTZIS, K.; ILIADIS, L.; ANEZAKIS, V. D. A deep spiking machine-hearing system for the case of invasive fish species. In: IEEE INTERNATIONAL CONFERENCE ON INNOVATIONS IN INTELLIGENT SYSTEMS AND APPLICATIONS (INISTA), 2017, Gdynia. *Proceedings…* Gdynia: IEEE, 2017. p. 23–28.

DESIDERÀ, E.; GUIDETTI, P.; PANZALIS, P.; NAVONE, A.; VALENTINI-POIRRIER, C. A.; BOISSERY, P.; GERVAISE, C. et al. Acoustic fish communities: sound diversity of rocky habitats reflects fish species diversity. *Marine Ecology Progress Series*, v. 608, p. 183–197, 2019.

DI IORIO, L.; AUDAX, M.; DETER, J.; HOLON, F.; LOSSENT, J.; GERVAISE, C.; BOISSERY, P. Biogeography of acoustic biodiversity of NW Mediterranean coralligenous reefs. *Scientific Reports*, v. 11, p. 1–10, 2021.

DI IORIO, L.; RAICK, X.; PARMENTIER, E.; BOISSERY, P.; VALENTINI-POIRIER, C. A.; GERVAISE, C. “Posidonia meadows calling”: a ubiquitous fish sound with monitoring potential. *Remote Sensing in Ecology and Conservation*, v. 4, p. 248–263, 2018.

DUARTE, C. M.; CHAPUIS, L.; COLLIN, S. P.; COSTA, D. P.; DEVASSY, R. P.; EGUILUZ, V. M.; ERBE, C. et al. The soundscape of the Anthropocene ocean. *Science*, v. 371, p. eaba4658, 2021.

ESFAHANIAN, M.; ZHUANG, H.; ERDOL, N. On contour-based classification of dolphin whistles by type. *Applied Acoustics*, v. 76, p. 274–279, 2014a.

ESFAHANIAN, M.; ZHUANG, H.; ERDOL, N. Sparse representation for classification of dolphin whistles by type. *The Journal of the Acoustical Society of America*, v. 136, p. EL1–EL7, 2014b.

ETTER, P. C. *Underwater acoustic modeling and simulation*. 3. ed. Boca Raton: CRC Press, 2003. 448 p.

FINE, M. L. Seasonal and geographical variation of the mating call of the oyster toadfish *Opsanus tau* L. *Oecologia*, v. 36, p. 45–57, 1978.

FINE, M. L.; LENHARDT, M. L. Shallow-water propagation of the toadfish mating call. *Comparative Biochemistry and Physiology Part A: Physiology*, v. 76, p. 225–231, 1983.

FISH, M. P.; MOWBRAY, W. H. *Sounds of Western North Atlantic fishes: a reference file of biological underwater sounds*. Baltimore: The Johns Hopkins Press, 1970.

FRASIER, K. E. A machine learning pipeline for classification of cetacean echolocation clicks in large underwater acoustic datasets. *PLoS Computational Biology*, v. 17, p. e1009613, 2021.

FRENCH, B.; WILSON, S.; HOLMES, T.; KENDRICK, A.; RULE, M.; RYAN, N. Comparing five methods for quantifying abundance and diversity of fish assemblages in seagrass habitat. *Ecological Indicators*, v. 124, p. 107415, 2021.

FROESE, R.; PAULY, D. (ed.). *FishBase*. Disponível em: <www.fishbase.org>. Versão: 02/2022. Acesso em: 19 jun. 2022.

GAIKWAD, S. K.; GAWALI, B. W.; YANNAWAR, P. A review on speech recognition technique. *International Journal of Computer Applications*, v. 10, p. 16–24, 2010.

GALAIDUK, R.; RADFORD, B. T.; WILSON, S. K.; HARVEY, E. S. Comparing two remote video survey methods for spatial predictions of the distribution and environmental niche suitability of demersal fishes. *Scientific Reports*, v. 7, p. 1–11, 2017.

GANNON, D. P. Passive acoustic techniques in fisheries science: a review and prospectus. *Transactions of the American Fisheries Society*, v. 137, p. 638–656, 2008.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. Cambridge, MA: MIT Press, 2016. 785 p.

GOODWIN, M.; HALVORSEN, K. T.; JIAO, L.; KNAUSGÅRD, K. M.; MARTIN, A. H.; MOYANO, M.; OOMEN, R. A. et al. Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. *ICES Journal of Marine Science*, v. 79, p. 319–336, 2022.

GUYOT, P.; ALIX, F.; GUERIN, T.; LAMBEAUX, E.; ROTUREAU, A. Fish migration monitoring from audio detection with CNNs. In: INTERNATIONAL AUDIO MOSTLY CONFERENCE, 16., 2021, Trento. *Proceedings…* Trento: ACM, 2021. p. 244–247.

HALPERN, B. S.; WALBRIDGE, S.; SELKOE, K. A.; KAPPEL, C. V.; MICHELI, F.; D'AGROSA, C.; BRUNO, J. F. et al. A global map of human impact on marine ecosystems. *Science*, v. 319, p. 948–952, 2008.

HARAKAWA, R.; OGAWA, T.; HASEYAMA, M.; AKAMATSU, T. Automatic detection of fish sounds based on multi-stage classification including logistic regression via adaptive feature weighting. *The Journal of the Acoustical Society of America*, v. 144, p. 2709–2718, 2018.

HARRIS, S. A.; SHEARS, N. T.; RADFORD, C. A. Ecoacoustic indices as proxies for biodiversity on temperate reefs. *Methods in Ecology and Evolution*, v. 7, p. 713–724, 2016.

HARVEY, E.; FLETCHER, D.; SHORTIS, M. R.; KENDRICK, G. A. A comparison of underwater visual distance estimates made by scuba divers and a stereo-video system: implications for underwater visual census of reef fish abundance. *Marine and Freshwater Research*, v. 55, p. 573–580, 2004.

HAWKINS, A. D.; POPPER, A. N. Effects of man-made sound on fishes. In: SLABEKORN, H.; DOOLING, R.; POPPER, A.; FAY, R. (ed.). *Effects of anthropogenic noise on animals*. 1. ed. New York: Springer, p. 145–177, 2018.

HELLMRICH, L. S.; SAUNDERS, B. J.; PARKER, J. R.; GOETZE, J. S.; HARVEY, E. S. Stereo-ROV surveys of tropical reef fishes are comparable to stereo-DOVs with reduced behavioural biases. *Estuarine, Coastal and Shelf Science*, v. 281, p. 108210, 2023.

HERRERA-BOYER, P.; PEETERS, G.; DUBNOV, S. Automatic classification of musical instrument sounds. *Journal of New Music Research*, v. 32, p. 3–21, 2003.

HILDEBRAND, J. A. Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series*, v. 395, p. 5–20, 2009.

HOLMLUND, C. M.; HAMMER, M. Ecosystem services generated by fish populations. *Ecological Economics*, v. 29, p. 253–268, 1999.

HOLT, D. E.; JOHNSTON, C. E. Evidence of the Lombard effect in fishes. *Behavioral Ecology*, v. 25, p. 819–826, 2014.

HUANG, X.; HU, Z.; LIN, L. Deep clustering based on embedded auto-encoder. *Soft Computing*, v. 27, p. 1075–1090, 2023.

HUANG, C. J.; YANG, Y. J.; YANG, D. X.; CHEN, Y. J. Frog classification using machine learning techniques. *Expert Systems with Applications*, v. 36, p. 3737–3743, 2009.

IBRAHIM, A. K.; CHÉRUBIN, L. M.; ZHUANG, H.; SCHÄRER-UMPIERRE, M. T.; DALGLEISH, F.; ERDOL, N.; OUYANG, B. et al. An approach for automatic classification of grouper vocalizations with passive acoustic monitoring. *The Journal of the Acoustical Society of America*, v. 143, p. 666–676, 2018a.

IBRAHIM, A. K.; ZHUANG, H.; CHÉRUBIN, L. M.; SCHÄRER-UMPIERRE, M. T.; ERDOL, N. Automatic classification of grouper species by their sounds using deep neural networks. *The Journal of the Acoustical Society of America*, v. 144, p. EL196–EL202, 2018b.

IBRAHIM, A. K.; ZHUANG, H.; CHÉRUBIN, L. M.; SCHÄRER-UMPIERRE, M. T.; NEMETH, R. S.; ERDOL, N.; ALI, A. M. Transfer learning for efficient classification of grouper sound. *The Journal of the Acoustical Society of America*, v. 148, p. EL260–EL266, 2020.

JABBAR, H.; KHAN, R. Z. Methods to avoid overfitting and underfitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, v. 70, p. 978–981, 2015.

JASIM, W. N.; SADDAM, S. A. W.; HARFASH, E. J. Wind sounds classification using different audio feature extraction techniques. *Informatica*, v. 45, p. 57–65, 2021.

JIANG, J. J.; BU, L. R.; DUAN, F. J.; WANG, X. Q.; LIU, W.; SUN, Z. B.; LI, C. Y. Whistle detection and classification for whales based on convolutional neural networks. *Applied Acoustics*, v. 150, p. 169–178, 2019.

JUBLIER, N.; BERTUCCI, F.; KÉVER, L.; COLLEYE, O.; BALLESTA, L.; NEMETH, R. S.; LECCHINI, D. et al. Passive monitoring of phenological acoustic patterns reveals the sound of the camouflage grouper, *Epinephelus polyphekadion*. *Aquatic Conservation: Marine and Freshwater Ecosystems*, v. 30, p. 42–52, 2020.

JUHEL, J. B.; MARQUES, V.; UTAMA, R. S.; VIMONO, I. B.; SUGEHA, H. Y.; KADARUSMAN, K.; COCHET, C. et al. Estimating the extended and hidden species diversity from environmental DNA in hyper-diverse regions. *Ecography*, v. 10, p. 06299, 2022.

KAHL, S.; WOOD, C. M.; EIBL, M.; KLINCK, H. BirdNET: a deep learning solution for avian diversity monitoring. *Ecological Informatics*, v. 61, p. 101236, 2021.

KAMAL, S.; MOHAMMED, S. K.; PILLAI, P. S.; SUPRIYA, M. H. Deep learning architectures for underwater target recognition. In: SYMPOSIUM ON OCEAN ELECTRONICS (SYMPOL), 2013, Kochi. *Proceedings…* Kochi: IEEE, 2013. p. 48–54.

KASUMYAN, A. O. Acoustic signaling in fish. *Journal of Ichthyology*, v. 49, p. 963–1020, 2009.

KIM, E. B.; FRASIER, K. E.; MCKENNA, M. F.; KOK, A. C.; PEAVEY REEVES, L. E.; OESTREICH, W. K.; ARRIETA, G. et al. SoundScape learning: an automatic method for separating fish chorus in marine soundscapes. *The Journal of the Acoustical Society of America*, v. 153, p. 1710–1722, 2023.

KOTTEGE, N.; JURDAK, R.; KROON, F.; JONES, D. Automated detection of broadband clicks of freshwater fish using spectro-temporal features. *The Journal of the Acoustical Society of America*, v. 137, p. 2502–2511, 2015.

KUMAR, C. S.; JOSE, J. A. A wavelet based time-frequency descriptor for automatic classification of acoustic signals of fishes. In: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING, INSTRUMENTATION AND CONTROL TECHNOLOGIES (ICICICT), 2., 2019, Kannur. *Proceedings…* Kannur: IEEE, 2019. p. 1059–1063.

LACOURSIÈRE-ROUSSEL, A.; ROSABAL, M.; BERNATCHEZ, L. Estimating fish abundance and biomass from eDNA concentrations: variability among capture methods and environmental conditions. *Molecular Ecology Resources*, v. 16, p. 1401–1414, 2016.

LADICH, F. Fish bioacoustics. *Current Opinion in Neurobiology*, v. 28, p. 121–127, 2014.

LADICH, F. Acoustic communication in fishes: temperature plays a role. *Fish and Fisheries*, v. 19, p. 598–612, 2018.

LADICH, F. Ecology of sound communication in fishes. *Fish and Fisheries*, v. 20, p. 552–563, 2019.

LADICH, F. Shut up or shout loudly: predation threat and sound production in fishes. *Fish and Fisheries*, v. 23, p. 227–238, 2022.

LAMY, T.; PITZ, K. J.; CHAVEZ, F. P.; YORKE, C. E.; MILLER, R. J. Environmental DNA reveals the fine-grained and hierarchical spatial structure of kelp forest fish communities. *Scientific Reports*, v. 11, p. 14439, 2021.

LAPLANTE, J. F.; AKHLOUFI, M. A.; GERVAISE, C. Deep learning for marine bioacoustics and fish classification using underwater sounds. In: CANADIAN CONFERENCE ON ELECTRICAL AND COMPUTER ENGINEERING (CCECE), 2022, Halifax. *Proceedings…* Halifax: IEEE, 2022. p. 288–293.

LASSECK, M. Acoustic bird detection with deep convolutional neural networks. In: DETECTION AND CLASSIFICATION OF ACOUSTIC SCENES AND EVENTS (DCASE), 2018. *Proceedings…* Berlin: Museum für Naturkunde, 2018. p. 143–147.

LEE, H.; PHAM, P.; LARGMAN, Y.; NG, A. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: NEURAL INFORMATION PROCESSING SYSTEMS (NIPS), 23., 2009, Vancouver. *Proceedings…* Vancouver, 2009.

LIN, T. H.; TSAO, Y.; WANG, Y. H.; YEN, H. W.; LU, S. S. Computing biodiversity change via a soundscape monitoring network. In: PACIFIC NEIGHBORHOOD CONSORTIUM ANNUAL CONFERENCE AND JOINT MEETINGS (PNC), 2017, Tainan. *Proceedings…* Tainan: IEEE, 2017. p. 128–133.

LIN, T. H.; TSAO, Y.; AKAMATSU, T. Comparison of passive acoustic soniferous fish monitoring with supervised and unsupervised approaches. *The Journal of the Acoustical Society of America*, v. 143, p. EL278–EL284, 2018.

LIN, T. H.; TSAO, Y. Source separation in ecoacoustics: a roadmap towards versatile soundscape information retrieval. *Remote Sensing in Ecology and Conservation*, v. 6, p. 236–247, 2020.

LINDSETH, A. V.; LOBEL, P. S. Underwater soundscape monitoring and fish bioacoustics: a review. *Fishes*, v. 3, p. 36, 2018.

LIU, Y.; NIU, H.; LI, Z. A multi-task learning convolutional neural network for source localization in deep ocean. *The Journal of the Acoustical Society of America*, v. 148, p. 873–883, 2020.

LOBEL, P. S.; MANN, D. A. Spawning sounds of the damselfish, *Dascyllus albisella* (Pomacentridae), and relationship to male size. *Bioacoustics*, v. 6, p. 187–198, 1995.

LOBEL, P. S. Diversity of fish spawning sounds and the application of passive acoustic monitoring. *Bioacoustics*, v. 12, p. 286–289, 2002.

LOOBY, A.; COX, K.; BRAVO, S.; ROUNTREE, R.; JUANES, F.; REYNOLDS, L. K.; MARTIN, C. W. A quantitative inventory of global soniferous fish diversity. *Reviews in Fish Biology and Fisheries*, v. 32, p. 581–595, 2022a.

LOOBY, A.; RIERA, A.; VELA, S.; COX, K.; BRAVO, S.; ROUNTREE, R.; JUANES, F. et al. *FishSounds*. Disponível em: http://www.fishsounds.net. Acesso em: 19 jun. 2022, 2022b.

LU, X.; TSAO, Y.; MATSUDA, S.; HORI, C. Speech enhancement based on deep denoising autoencoder. In: INTERSPEECH, 2013, Lyon. *Proceedings…* Lyon: ISCA, 2013. p. 436–440.

MCLOUGHLIN, M. P.; STEWART, R.; MCELLIGOTT, A. G. Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface*, v. 16, p. 20190225, 2019.

MAHALE, V. P.; CHANDA, K.; CHAKRABORTY, B.; SALKAR, T.; SREEKANTH, G. B. Biodiversity assessment using passive acoustic recordings from off-reef location – unsupervised learning to classify fish vocalization. *The Journal of the Acoustical Society of America*, v. 153, p. 1534–1553, 2023.

MALFANTE, M.; MARS, J. I.; DALLA MURA, M.; GERVAISE, C. Automatic fish sounds classification. *The Journal of the Acoustical Society of America*, v. 143, p. 2834–2846, 2018a.

MALFANTE, M.; MOHAMMED, O.; GERVAISE, C.; DALLA MURA, M.; MARS, J. I. Use of deep features for the automatic classification of fish sounds. In: OCEANS – MTS/IEEE KOBE TECHNO-OCEANS, 2018, Kobe. *Proceedings…* Kobe: IEEE, 2018b. p. 1–5.

MALLET, D.; PELLETIER, D. Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). *Fisheries Research*, v. 154, p. 44–62, 2014.

MANCUSI, M.; ZONCA, N.; RODOLÀ, E.; ZUFFI, S. Fish sounds: towards the evaluation of marine acoustic biodiversity through data-driven audio source separation. 2022. Disponível em: https://doi.org/10.48550/arXiv.2201.05013. Acesso em: 14 jan. 2022. Preprint.

MANN, D. A. Remote sensing of fish using passive acoustic monitoring. *Acoustics Today*, v. 8, p. 8–15, 2012.

MANN, D. A.; HAWKINS, A. D.; JECH, J. M. Active and passive acoustics to locate and study fish. In: WEBB, J.; POPPER, A.; FAY, R. (ed.). *Fish bioacoustics*. New York: Springer, p. 279–309, 2008.

MANN, D. A.; LOBEL, P. S. Acoustic behavior of the damselfish *Dascyllus albisella*: behavioral and geographic variation. *Environmental Biology of Fishes*, v. 51, p. 421–428, 1998.

MEHYADIN, A. E.; ABDULAZEEZ, A. M.; HASAN, D. A.; SAEED, J. N. Birds sound classification based on machine learning algorithms. *Asian Journal of Research in Computer Science*, v. 9, p. 1–11, 2021.

MELLINGER, D. K.; STAFFORD, K. M.; MOORE, S. E.; DZIAK, R. P.; MATSUMOTO, H. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography*, v. 20, p. 36–45, 2007.

MERCHANT, N. D.; FRISTRUP, K. M.; JOHNSON, M. P.; TYACK, P. L.; WITT, M. J.; BLONDEL, P.; PARKS, S. E. Measuring acoustic habitats. *Methods in Ecology and Evolution*, v. 6, p. 257–265, 2015.

MIKSIS-OLDS, J. L.; MARTIN, B.; TYACK, P. L. Exploring the ocean through soundscapes. *Acoustics Today*, v. 14, p. 26–34, 2018.

MILLER, B. S.; MADHUSUDHANA, S.; AULICH, M. G.; KELLY, N. Deep learning algorithm outperforms experienced human observer at detection of blue whale D-calls: a double-observer analysis. *Remote Sensing in Ecology and Conservation*, v. 9, p. 104–116, 2023.

MINELLO, M.; CALADO, L.; XAVIER, F. C. Ecoacoustic indices in marine ecosystems: a review on recent developments, challenges, and future directions. *ICES Journal of Marine Science*, v. 78, p. 3066–3074, 2021.

MISHACHANDAR, B.; VAIRAMUTHU, S. Diverse ocean noise classification using deep learning. *Applied Acoustics*, v. 181, p. 108141, 2021.

MONCZAK, A.; JI, Y.; SOUEIDAN, J.; MONTIE, E. W. Automatic detection, classification, and quantification of sciaenid fish calls in an estuarine soundscape in the southeast United States. *PLoS One*, v. 14, p. e0209914, 2019.

MOONEY, T. A.; DI IORIO, L.; LAMMERS, M.; LIN, T. H.; NEDELEC, S. L.; PARSONS, M.; RADFORD, C. et al. Listening forward: approaching marine biodiversity assessments using acoustic methods. *Royal Society Open Science*, v. 7, p. 201287, 2020.

MUNDAY, P. L.; WILSON, S. K. Comparative efficacy of clove oil and other chemicals in anaesthetization of *Pomacentrus amboinensis*, a coral reef fish. *Journal of Fish Biology*, v. 51, p. 931–938, 1997.

MUNGER, J. E.; HERRERA, D. P.; HAVER, S. M.; WATERHOUSE, L.; MCKENNA, M. F.; DZIAK, R. P.; GEDAMKE, J. et al. Machine learning analysis reveals relationship between pomacentrid calls and environmental cues. *Marine Ecology Progress Series*, v. 681, p. 197–210, 2022.

MOUY, X.; ROUNTREE, R.; JUANES, F.; DOSSO, S. E. Cataloging fish sounds in the wild using combined acoustic and video recordings. *The Journal of the Acoustical Society of America*, v. 143, p. EL333–EL339, 2018.

MOUY, X.; BLACK, M.; COX, K.; QUALLEY, J.; DOSSO, S.; JUANES, F. Identification of fish sounds in the wild using a set of portable audio–video arrays. *Methods in Ecology and Evolution*, v. 00, p. 1–22, 2023.

MURPHY, H. M.; JENKINS, G. P. Observational methods used in marine spatial monitoring of fishes and associated habitats: a review. *Marine and Freshwater Research*, v. 61, p. 236–252, 2010.

MYRBERG, A. A.; HA, S. J.; SHAMBLOTT, M. J. The sounds of bicolor damselfish (*Pomacentrus partitus*): predictors of body size and a spectral basis for individual recognition and assessment. *The Journal of the Acoustical Society of America*, v. 94, p. 3067–3070, 1993.

NANNI, L.; MAGUOLO, G.; PACI, M. Data augmentation approaches for improving animal audio classification. *Ecological Informatics*, v. 57, p. 101084, 2020.

NIU, H.; REEVES, E.; GERSTOFT, P. Source localization in an ocean waveguide using supervised machine learning. *The Journal of the Acoustical Society of America*, v. 142, p. 1176–1188, 2017.

NODA, J. J.; TRAVIESO, C. M.; SÁNCHEZ-RODRÍGUEZ, D. Automatic taxonomic classification of fish based on their acoustic signals. *Applied Sciences*, v. 6, p. 443, 2016.

OLIVER, T. H.; HEARD, M. S.; ISAAC, N. J.; ROY, D. B.; PROCTER, D.; EIGENBROD, F.; FRECKLETON, R. et al. Biodiversity and resilience of ecosystem functions. *Trends in Ecology & Evolution*, v. 30, p. 673–684, 2015.

OSGOOD, G. J.; MCCORD, M. E.; BAUM, J. K. Using baited remote underwater videos (BRUVs) to characterize chondrichthyan communities in a global biodiversity hotspot. *PLoS One*, v. 14, p. 0225859, 2019.

OZANICH, E.; THODE, A.; GERSTOFT, P.; FREEMAN, L. A.; FREEMAN, S. Deep embedded clustering of coral reef bioacoustics. *The Journal of the Acoustical Society of America*, v. 149, p. 2587–2601, 2021.

PADOVESE, B.; FRAZAO, F.; KIRSEBOM, O. S.; MATWIN, S. Data augmentation for the classification of North Atlantic right whales upcalls. *The Journal of the Acoustical Society of America*, v. 149, p. 2520–2530, 2021.

PANDEYA, Y. R.; LEE, J. Domestic cat sound classification using transfer learning. *International Journal of Fuzzy Logic and Intelligent Systems*, v. 18, p. 154–160, 2018.

PARMENTIER, E.; BERTUCCI, F.; BOLGAN, M.; LECCHINI, D. How many fish could be vocal? An estimation from a coral reef (Moorea Island). *Belgian Journal of Zoology*, v. 151, p. 82, 2021.

PARMENTIER, E.; LAGARDÈRE, J. P.; VANDEWALLE, P.; FINE, M. L. Geographical variation in sound production in the anemonefish *Amphiprion akallopisos*. *Proceedings of the Royal Society B: Biological Sciences*, v. 272, p. 1697–1703, 2005.

PARSONS, M. J.; LIN, T. H.; MOONEY, T. A.; ERBE, C.; JUANES, F.; LAMMERS, M.; LI, S. et al. Sounding the call for a global library of underwater biological sounds. *Frontiers in Ecology and Evolution*, v. 10, p. 39, 2022.

PARSONS, S.; JONES, G. Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks. *Journal of Experimental Biology*, v. 203, p. 2641–2656, 2000.

PECK, M.; TAPILATU, R. F.; KURNIATI, E.; ROSADO, C. Rapid coral reef assessment using 3D modelling and acoustics: acoustic indices correlate to fish abundance, diversity and environmental indicators in West Papua, Indonesia. *PeerJ*, v. 9, p. e10761, 2021.

PHILLIPS, C. T.; JOHNSTON, C. E. Geographical divergence of acoustic signals in *Cyprinella galactura*, the whitetail shiner (Cyprinidae). *Animal Behaviour*, v. 75, p. 617–626, 2008.

PICHLER, M.; HARTIG, F. Machine learning and deep learning — a review for ecologists. *Methods in Ecology and Evolution*, v. 14, p. 994–1016, 2023.

PIERETTI, N.; DANOVARO, R. Acoustic indexes for marine biodiversity trends and ecosystem health. *Philosophical Transactions of the Royal Society B*, v. 375, p. 20190447, 2020.

PUTLAND, R. L.; MACKIEWICZ, A. G.; MENSINGER, A. F. Localizing individual soniferous fish using passive acoustic monitoring. *Ecological Informatics*, v. 48, p. 60–68, 2018.

RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, v. 11, p. 185, 2019.

RICCARDI, G.; HAKKANI-TUR, D. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, v. 13, p. 504–511, 2005.

RICE, A. N.; FARINA, S. C.; MAKOWSKI, A. J.; KAATZ, I. M.; LOBEL, P. S.; BEMIS, W. E.; BASS, A. H. Evolutionary patterns in sound production across fishes. *Ichthyology & Herpetology*, v. 110, p. 1–12, 2022.

RICE, A. N.; SOLDEVILLA, M. S.; QUINLAN, J. A. Nocturnal patterns in fish chorusing off the coasts of Georgia and eastern Florida. *Bulletin of Marine Science*, v. 93, p. 455–474, 2017.

ROBERTSON, D. R.; SMITH-VANIZ, W. F. Use of clove oil in collecting coral reef fishes for research. *Marine Ecology Progress Series*, v. 401, p. 295–302, 2010.

ROCH, M. A.; GERSTOFT, P.; KOSTEK, B.; MICHALOPOULOU, Z. H. How machine learning contributes to solve acoustical problems. *The Journal of the Acoustical Society of America*, v. 17, p. 48–57, 2021.

ROLIM, F. A.; RODRIGUES, P. F.; LANGLOIS, T.; NEVES, L. M.; GADIG, O. B. A comparison of stereo-videos and visual census methods for assessing subtropical rocky reef fish assemblage. *Environmental Biology of Fishes*, v. 105, p. 413–429, 2022.

ROUNTREE, R. A.; GILMORE, R. G.; GOUDEY, C. A.; HAWKINS, A. D.; LUCZKOVICH, J. J.; MANN, D. A. Listening to fish: applications of passive acoustics to fisheries science. *Fisheries*, v. 31, p. 433–446, 2006.

ROWELL, T. J.; DEMER, D. A.; ABURTO-OROPEZA, O.; COTA-NIETO, J. J.; HYDE, J. R.; ERISMAN, B. E. Estimating fish abundance at spawning aggregations from courtship sound levels. *Scientific Reports*, v. 7, p. 1–14, 2017.

ROWELL, T. J.; SCHÄRER, M. T.; APPELDOORN, R. S.; NEMETH, M. I.; MANN, D. A.; RIVERA, J. A. Sound production as an indicator of red hind density at a spawning aggregation. *Marine Ecology Progress Series*, v. 462, p. 241–250, 2012.

RUIZ-BLAIS, S.; RIVERA-CHAVARRIA, M. R.; CAMACHO, A. Autonomous detection of neotropical sciaenid fishes. *Proceedings of Meetings on Acoustics*, v. 18, p. 010001, 2012.

RUPPÉ, L.; CLÉMENT, G.; HERREL, A.; BALLESTA, L.; DÉCAMPS, T.; KÉVER, L.; PARMENTIER, E. Environmental constraints drive the partitioning of the soundscape in fishes. *Proceedings of the National Academy of Sciences*, v. 112, p. 6092–6097, 2015.

SALEH, A.; SHEAVES, M.; RAHIMI AZGHADI, M. Computer vision and deep learning for fish classification in underwater habitats: a survey. *Fish and Fisheries*, v. 23, p. 977–999, 2022.

SARANGI, S. K.; SAHA, G. A novel approach in feature level for robust text-independent speaker identification system. In: INTERNATIONAL CONFERENCE ON INTELLIGENT HUMAN COMPUTER INTERACTION (IHCI), 4., 2012, Kharagpur. *Proceedings…* Kharagpur: IEEE, 2012. p. 1–5.

SATTAR, F.; CULLIS-SUZUKI, S.; JIN, F. Identification of fish vocalizations from ocean acoustic data. *Applied Acoustics*, v. 110, p. 248–255, 2016.

SCHÄRER, M. T.; NEMETH, M. I.; MANN, D.; LOCASCIO, J.; APPELDOORN, R. S.; ROWELL, T. J. Sound production and reproductive behavior of yellowfin grouper, *Mycteroperca venenosa* (Serranidae) at a spawning aggregation. *Copeia*, v. 2012, p. 135–144, 2012.

SHARMA, G.; UMAPATHY, K.; KRISHNAN, S. Trends in audio signal feature extraction methods. *Applied Acoustics*, v. 158, p. 107020, 2020.

SHIU, Y.; PALMER, K. J.; ROCH, M. A.; FLEISHMAN, E.; LIU, X.; NOSAL, E. M.; HELBLE, T. et al. Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, v. 10, p. 607, 2020.

SILVA, D. F.; DE SOUZA, V. M.; BATISTA, G. E.; KEOGH, E.; ELLIS, D. P. Applying machine learning and audio analysis techniques to insect recognition in intelligent traps. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS (ICMLA), 12., 2013. *Proceedings…* 2013. v. 1, p. 99–104.

SMARAGDAKIS, C.; TAROUDAKIS, M. I. Acoustic signal characterization based on hidden Markov models with applications to geoacoustic inversions. *The Journal of the Acoustical Society of America*, v. 148, p. 2337–2350, 2020.

STOWELL, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, v. 10, p. e13152, 2022.

STOWELL, D.; PETRUSKOVÁ, T.; ŠÁLEK, M.; LINHART, P. Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *Journal of the Royal Society Interface*, v. 16, p. 20180940, 2019a.

STOWELL, D.; WOOD, M. D.; PAMUŁA, H.; STYLIANOU, Y.; GLOTIN, H. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, v. 10, p. 368–380, 2019b.

SUN, Y. J.; YEN, S. C.; LIN, T. H. Soundscape_IR: a source separation toolbox for exploring acoustic diversity in soundscapes. *Methods in Ecology and Evolution*, v. 13, p. 2347–2355, 2022.

TAMPOSIS, I. A.; TSIRIGOS, K. D.; THEODOROPOULOU, M. C.; KONTOU, P. I.; BAGOS, P. G. Semi-supervised learning of hidden Markov models for biological sequence analysis. *Bioinformatics*, v. 35, p. 2208–2215, 2019.

TAVOLGA, W. N. Sound production and detection. In: HOAR, W.; RANDALL, D. (ed.). *Fish physiology*. v. 5. London: Academic Press, 1971. p. 135–205.

TELLECHEA, J. S.; MARTINEZ, C.; FINE, M. L.; NORBIS, W. Sound production in the whitemouth croaker and relationship between fish size and disturbance call characteristics. *Environmental Biology of Fishes*, v. 89, p. 163–172, 2010.

TOLKOVA, I.; KLINCK, H. Source separation with an acoustic vector sensor for terrestrial bioacoustics. *The Journal of the Acoustical Society of America*, v. 152, p. 1123–1134, 2022.

TUIA, D.; KELLENBERGER, B.; BEERY, S.; COSTELLOE, B. R.; ZUFFI, S.; RISSE, B.; MATHIS, A. et al. Perspectives in machine learning for wildlife conservation. *Nature Communications*, v. 13, p. 792, 2022.

USMAN, A. M.; OGUNDILE, O. O.; VERSFELD, D. J. Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access*, v. 8, p. 105181–105206, 2020.

VAN PARIJS, S. M.; CLARK, C. W.; SOUSA-LIMA, R. S.; PARKS, S. E.; RANKIN, S.; RISCH, D.; VAN OPZEELAND, I. C. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. *Marine Ecology Progress Series*, v. 395, p. 21–36, 2009.

VERZIJDEN, M. N.; VAN HEUSDEN, J.; BOUTON, N.; WITTE, F.; TEN CATE, C.; SLABBEKOORN, H. Sounds of male Lake Victoria cichlids vary within and between species and affect female mate preferences. *Behavioral Ecology*, v. 21, p. 548–555, 2010.

VIEIRA, M.; FONSECA, P. J.; AMORIM, M. C. P.; TEIXEIRA, C. J. 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*, v. 138, p. 3941–3950, 2015.

VIEIRA, M.; PEREIRA, B. P.; POUSÃO-FERREIRA, P.; FONSECA, P. J.; AMORIM, M. C. P. Seasonal variation of captive meagre acoustic signalling: a manual and automatic recognition approach. *Fishes*, v. 4, p. 28, 2019.

WADDELL, E. E.; RASMUSSEN, J. H.; ŠIROVIC, A. Applying artificial intelligence methods to detect and classify fish calls from the Northern Gulf of Mexico. *Journal of Marine Science and Engineering*, v. 9, p. 1128, 2021.

WALL, C. C.; SIMARD, P.; LEMBKE, C.; MANN, D. A. Large-scale passive acoustic monitoring of fish sound production on the West Florida Shelf. *Marine Ecology Progress Series*, v. 484, p. 173–188, 2013.

WANG, J. C.; WANG, J. F.; HE, K. W.; HSU, C. S. Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), 2006, Vancouver. *Proceedings…* Vancouver: IEEE, 2006. p. 1731–1735.

WATSON, D. L.; HARVEY, E. S.; ANDERSON, M. J.; KENDRICK, G. A. A comparison of temperate reef fish assemblages recorded by three underwater stereo-video techniques. *Marine Biology*, v. 148, p. 415–425, 2005.

WILLIS, T. J. Visual census methods underestimate density and diversity of cryptic reef fishes. *Journal of Fish Biology*, v. 59, p. 1408–1411, 2001.

WILSON, K. C.; SEMMENS, B. X.; GITTINGS, S. R.; MCCOY, C.; PATTENGILL-SEMMENS, C. V.; ŠIROVIC, A. Grouper source levels and aggregation dynamics inferred from passive acoustic localization at a multispecies spawning site. *The Journal of the Acoustical Society of America*, v. 151, p. 3052–3065, 2022.

ZHANG, L.; WANG, D.; BAO, C.; WANG, Y.; XU, K. Large-scale whale call classification by transfer learning on multi-scale waveforms and time-frequency features. *Applied Sciences*, v. 9, p. 1020, 2019.

ZHAO, H.; HUANG, X.; LIU, W.; YANG, L. Environmental sound classification based on feature fusion. *MATEC Web of Conferences*, v. 173, p. 03059, 2018.

ZHUANG, F.; QI, Z.; DUAN, K.; XI, D.; ZHU, Y.; ZHU, H.; XIONG, H. et al. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, v. 109, p. 43–76, 2020.

3. CAPÍTULO 2 - OPTIMAL FEATURE SELECTION AND MODEL EXPLANATION FOR REEF FISH SOUND CLASSIFICATION

Published in Philosophical Transactions of the Royal Society B, v.380, n.1928 (June, 2025). DOI: https://doi.org/10.1098/rstb.2024.0055.

ABSTRACT

Fish produce a wide variety of sounds that contribute to the soundscapes of aquatic environments. In reef systems, these sounds are important acoustic cues for various ecological processes. Artificial intelligence methods to detect, classify and identify fish sounds have become increasingly common. This study proposes the classification of unknown fish sounds recorded in a subtropical rocky reef using different feature sets, data augmentation and explainable artificial intelligence tools. We used different supervised algorithms (naive Bayes, random forest, decision trees and multilayer perceptron) to perform a multiclass classification of four classes of fish pulsed sounds. The proposed models showed excellent performances, achieving 98.1% of correct classification with multilayer perceptron using data augmentation. Explainable artificial intelligence allowed us to identify which features contributed to predict each sound class. Recognizing and characterizing these sounds is key to better understanding diel behaviours and functional roles associated with critical reef ecological processes. This article is part of the theme issue 'Acoustic monitoring for tropical ecology and conservation'.

1. INTRODUCTION

Fish produce a wide variety of sounds, contributing to the soundscapes of aquatic environments. Communication sounds are used in different behavioural contexts, such as reproduction (courtship and mating), aggregation, feeding and agonistic interactions (KASUMYAN, 2009; LADICH, 2019). Sounds produced by fish have distinct spectral and temporal features, with many species producing pulsed and repetitive sounds (WALL et al., 2013; PUTLAND et al., 2018). These features can enable species identification, behavioural patterns and reproductive and spawning conditions (BOLGAN et al., 2020). In reef systems, sounds are important acoustic cues for different ecological processes, including larvae orientation and settlement and habitat use (LOBEL et al., 2010).

A common approach to study fish sounds is Passive Acoustic Monitoring (PAM), which allows continuous and long-term assessment of soundscapes (MCWILLIAM et al., 2017; LIMA et al., 2024). However, these long-term recordings are producing an increasing amount of data, which makes manual analysis unfeasible and requires robust computational methods (BIANCO et al., 2019). Therefore, the use of artificial intelligence (AI) to detect, classify and identify fish sounds is becoming increasingly common (BARROSO et al., 2023). Although machine learning and deep learning have gained attention in this field in recent years, the interpretability of the models remains a gap.

One of the most challenging tasks concerning the classification problem using learning-based approaches is to explain the predictions generated by the models. The algorithms often operate like ‘black boxes’, and it is difficult to understand how the models achieve such performance, and which parameters were responsible for predicting each class (MOLNAR, 2019). Explainable Artificial Intelligence (XAI) has received increasing attention in various fields such as medicine (HULSEN, 2023) remote sensing (ISHIKAWA et al, 2023) soundscape ecology (PARCERISAS et al., 2023) and many others. This is a method inspired in game theory that provides explanations for machine learning models by computing the contribution of each feature to the prediction (MOLNAR, 2019). This technique allows us to understand which features are important and how they specifically impact individual predictions of the model. Therefore, XAI provides the interpretability of the models aligned to the specificities of the studied phenomenon/problem.

Feature selection is a fundamental step in sound classification using machine learning (SHARMA et al., 2020) as it helps to reduce dimensionality and influence the accuracy of the model (ABAYOMI-ALLI et al, 2022; JASIM et al., 2022), and it must reflect the specificities of the problem domain. Thus, understanding which features are more effective to represent fish sounds can significantly improve the detection, classification and identification tasks. However, the study of feature selection is still in its early stages in the field of fish sounds, and it is possible that currently underrated features could be good descriptors of these sounds. Timbral texture features (TTFs), including the Mel Frequency Cepstral Coefficients (MFCCs), are widely used in music and human speech recognition. MFCCs provide a robust representation of sound structures (CHÉRUBIN et al., 2020) and have also been used in recent studies to detect and classify fish sounds (NODA et al., 2016; MALFANTE et al., 2018; HARAKAWA et al., 2018). However, few studies have demonstrated the individual contribution of each coefficient in predicting different sounds.

In this paper, we classified unknown fish sounds from subtropical rocky reefs in Arraial do Cabo (23°44'S−42°W), Brazil. Four supervised algorithms, naive Bayes (NB) (LEWIS, 1998), random forest (RF) (BREIMAN, 2001), decision trees (DT) (BREIMAN et al., 2017) and multilayer perceptron (MLP) (RUMELHART; MCCLELLAND, 1987) were used to classify four classes of fish pulsed sounds. Different feature sets were used, where the contribution of features for each class and the SHapley Additive exPlanations (SHAP) values were computed to understand the influence of the features on the classifiers. To the best of our knowledge, this is one of the first studies to use XAI to explain the classification model and assess the relationships between the features, allowing us to understand which features better describe each class of reef fish sounds.

2. MATERIAL AND METHODS

2.1. Data acquisition

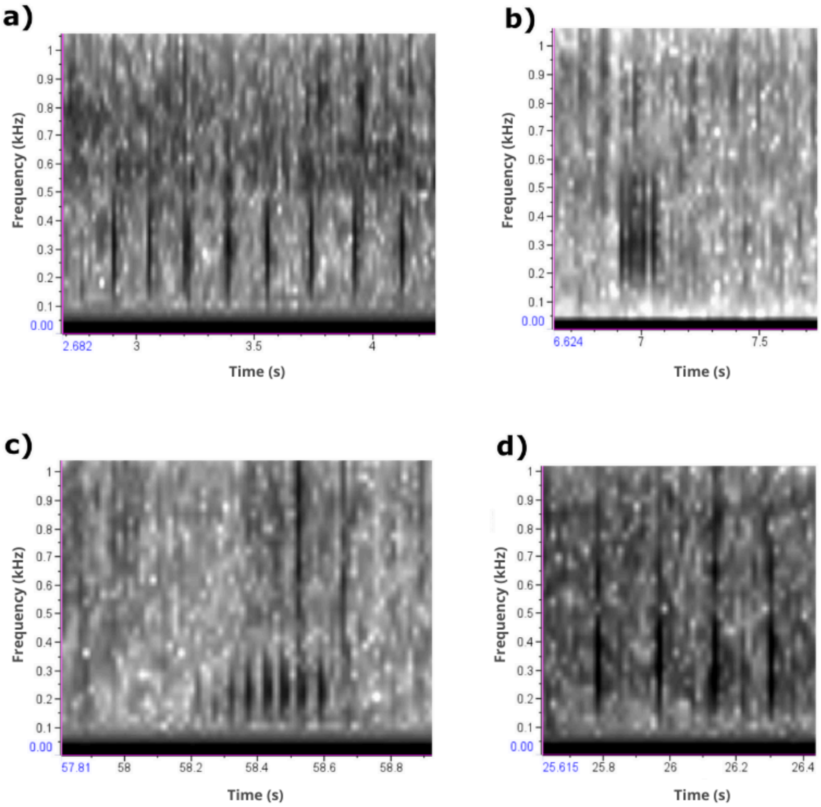
We used acoustic data from the long-term monitoring of the BIOCOM project (JESUS et al., 2020). The study was conducted on the island of Cabo Frio, Arraial do Cabo (23°44'S−42°W), Rio de Janeiro, south-eastern Brazil. The region is characterized by extensive subtropical rocky shores influenced by seasonal upwelling events, which imparts unique characteristics of a clash between tropical and warm-temperate biodiversity (CORDEIRO et al., 2014). Local upwelling brings intermediate deep waters with temperatures below 18° to shallow habitats, enriching trophic webs and promoting significant local hydrobiological variability (CARRIERE et al., 2009; VALENTIN, 1984). The combination of the upwelling regime and the diversity of reef habitats favours a rich fish community, with tropical and subtropical components (FERREIRA et al., 2001; CORDEIRO et al., 2016).

Acoustic data were collected using a recording system consisting of a stainless-steel pyramid structure with a 4-channel hydrophone (Hyd TP-1, Marsensing Ltda) deployed at a depth of 7.55 m and approximately 5 m from the local rocky shore. The recorder was configured with a sampling frequency of 52 734 Hz, 24-bit resolution and a duty cycle of 20% (i.e. a 1 min recording every 5 min). The sensitivity of the hydrophone was −174.9 dB re 1 μPa, with a flat response between 0.1 and 26 367 Hz. (For a brief description of the local soundscape, see electronic supplementary material, 'Methods'.) (JESUS et al., 2020; CAMPBELL et al., 2019).

2.2. Exploratory data analysis and preprocessing

To classify fish sounds, we selected sound files during different seasons of the year from the night, dusk and dawn periods, as these are known for fish acoustic activity (MANN; GROTHUES, 2009; RICEet al., 2017; SIDDAGANGAIAH et al., 2021). We grouped the sounds into four classes (figure 1) that were most common in the dataset. The classes are low-frequency sounds consisting of sequences of pulses: Class 1 is a pulse train with 3 to 9 pulses, call duration between 177 and 1387 ms, irregular inter-pulse intervals (IPIs) and a frequency range of 100−500 Hz; Class 2 is a pulse train with 4 pulses, call duration between 157 and 347 ms, almost regular IPIs and a frequency range of 100−600 Hz; Class 3 is a pulse train with 4 to 13 pulses, call duration between 195 and 910 ms, irregular IPIs and a frequency range of 100−400 Hz; and Class 4 is a pulse train with 3 to 7 pulses, with call duration between 216 and 939 ms, irregular IPIs and a frequency range of 100−700 Hz. (For more details regarding the sound classes see electronic supplementary material, 'Methods') (LOOBY, 2023; FROESE; PAULY, 2021; FISH; MOWBRAY, 1970; AMORIM, 2006; LADICH, 2014; CARRIÇO et al., 2019; DESIDERÀ et al., 2019).

FIGURA 1 - Spectrograms of the four classes of reef fish sounds. The four classes are low-frequency sounds, consisting of sequences of pulses. (a) Class 1, (b) Class 2, (c) Class 3, (d) Class 4.



Prior to the feature extraction step, the signals were pre-processed. Since the target sounds were of low frequency, filters were applied. For feature set 1, the raw sound files were filtered using the Raven Pro v. 1.6 bandpass filter tool with a lower limit = 100 Hz and an upper limit = 1000 Hz. For feature set 2, we developed a Python code to detect the target sounds according to their start and end times, previously extracted in feature set 1. In this step, the signals were downsampled (from 52 734 Hz to 5273 Hz) and filtered by frequency using an elliptic (Cauer) bandpass filter from the SciPy library. The filter was configured with a maximum ripple of 1 dB, a minimum attenuation of 20 dB and an optimal filter order, based on frequency limits for each signal.

2.3. Feature extraction

With the aim of identifying an optimal feature set to classify the signals, we proposed three different feature sets, totalling 40 features from different domains (time, frequency, amplitude and cepstral domain). The first set has a total of nine features, consisting of spectral features (high frequency, low frequency, peak frequency and centre frequency), temporal features (call duration, pulse duration, interpulse interval and number of pulses) and energy. To build this set, we manually extracted features by inspecting both oscillograms and spectrograms using Raven Pro v. 1.6 (window size = 1768 fft, overlap = 50%, colour scale = 'Greyscale'). Energy and spectral parameters were extracted directly using Raven functions, while temporal parameters were extracted by using various selections in the signals and extracting 'delta time' (CHARIF et al., 2010).

To explore TTFs, we used the librosa Python package (MCFEE et al., 2015), which has specialized functions for this purpose. To do this, we developed a code to detect fish signals (previously manually selected in the spectrograms). Leveraging librosa functions, we also extracted tempo (beats) to obtain possible information about the underlying rhythm of these signals. For this set, we selected 31 features (spectral centroid, spectral rolloff, zero crossing rate, spectral flux, tempo and 26 MFCCs) (ZELKOWITZ, 2010; GIANNAKOPOULOS; PIKRAKIS, 2014; TORRES-GARCÍA et al., 2021). The novelty of this approach is to evaluate the contribution of the MFCCs by interpreting the coefficients that were more important for each class. MFCCs provide information regarding energy distribution on the spectral envelope of signals (ZELKOWITZ, 2010). Each coefficient corresponds to a specific frequency band, but in a perceptual scale (the Mel scale). Thus, MFCC1 can be interpreted as the total energy obtained, MFCC2 as the weighted ratio of the energy of a low to a high frequency band, MFCC3 as the ratio of mid to low + high frequencies, MFCC4 as the ratio of two specific bands to two higher frequency bands, and so on (TRACEY et al., 2023). It is worth mentioning that the Mel scale provides higher resolution in the lower frequency bands, as it has an almost linear relationship with the Hertz scale below 1 KHz (RAO; MANJUNATH, 2017) .

Feature set 3 is a combination of sets 1 and 2, with 40 features. The description of the features used in each set is shown in electronic supplementary material, table S1. In addition to the three main feature sets, we tested other combinations of features by domain (cepstral, spectral and temporal) to investigate their contributions. These sets were (i) MFCCS only (n = 26), (ii) spectral features only (n = 8), (iii) temporal features only (n = 5), (iv) TTFs only (n = 30) (electronic supplementary material, table S2).

2.4. Data augmentation

Data augmentation is a technique commonly used to train machine learning algorithms on small and/or unbalanced datasets (STOWELL, 2022). Considering this, we selected only a few samples of fish calls to test the contribution of data augmentation applied to tabular data to the performance of the classifiers. Therefore, a small dataset of fish sounds was used to assess the effectiveness of data augmentation on tabular data. This is important because in many situations larger datasets of fish sounds are not available. The original datasets contained 120 samples (30 samples for each class). We augmented each dataset by using the standard deviations limits (summing and subtracting of each sample). After augmentation, the datasets had 360 samples (90 samples for each class). (For more details of the data augmentation process, see electronic supplementary material, ‘Methods’.)

2.5. Classification model

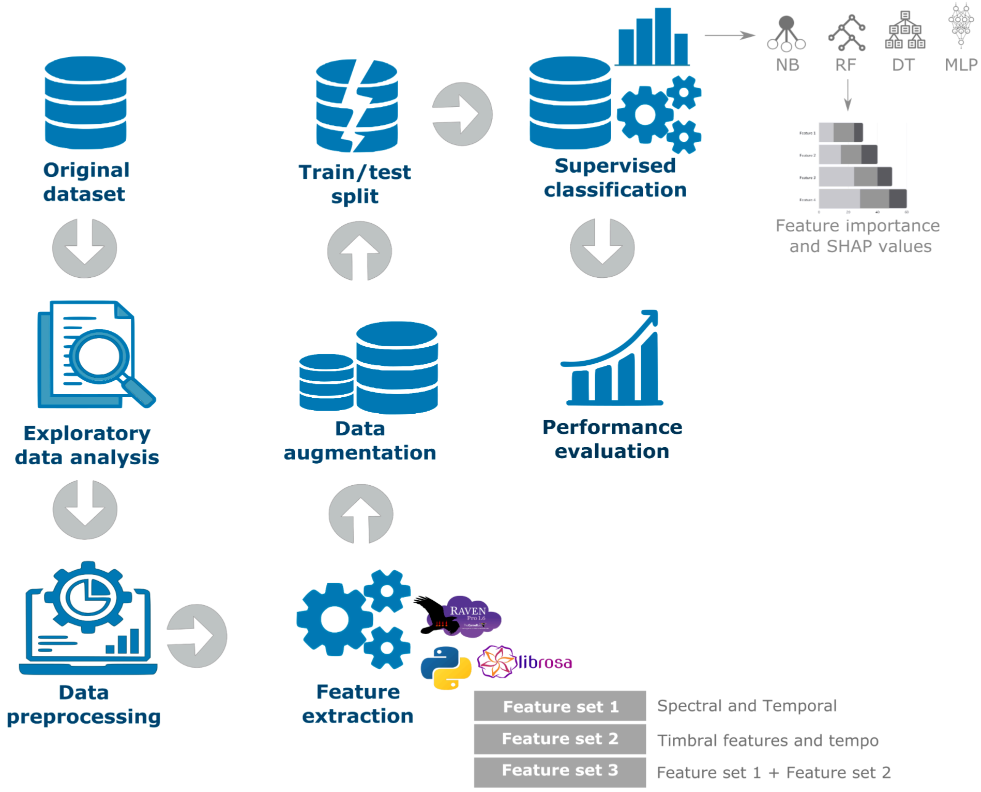
After the feature extraction process, fish sounds were classified using three supervised algorithms (NB, RF and DT) and a supervised neural network (MLP). With the aim of removing less important features (dimensionality reduction), we generated a random variable with a uniform distribution and added it to the feature vector to train an RF estimator to assess feature importance prior to the classification process. However, no feature was ranked below the random variable, and we assumed that all features were important. Tabular data of each feature set were used as input for these four algorithms. The classifiers were implemented using the Scikit-Learn Python package (PEDREGOSA, 2011). To train the supervised algorithms, the dataset was randomly split into 70% for training and 30% for testing. The supervised algorithms were run under Scikit-learn default parameters for all classifiers. The number of estimators for RF was set to 100. The MLP architecture developed for this work consisted of four hidden layers. The selected activation function was the Rectified Linear Unit (ReLU). The selected solver for weight optimization during the training step was the Adaptive Moment Estimation (Adam) algorithm. The regularization parameter (alpha), which helps to prevent overfitting, was set to 0.1. Training was performed over 200 epochs with a batch size of 8. For MLP, the dataset was also divided into 70% for training and 30% for testing.

Four metrics were used to evaluate the performance of the classification algorithms: accuracy, precision, recall and f1-score (MALFANTE et al., 2018; LAPLANTE et al., 2022). Accuracy is the ratio of the number of correct predictions to the total number of predictions, and it measures the overall correctness of a classification model; precision is calculated as the ratio of true positives to the total number of positive predictions made by the classifier, and it measures the correctness of positive predictions; recall is the ratio of the number of true positive predictions to the number of positive observations in the data; f1-score is the weighted average of precision and recall (SOKOLOVA; LAPALME, 2009; CAI et al., 2022). We also used a k-fold cross-validation method (k = 5) to evaluate the robustness of the classification model and the data augmentation effects. These results are presented in electronic supplementary material, table S3.

2.6. Model explanation

To provide a comprehensive explanation of the proposed classifiers, we used SHAP. We used the shapExplainer algorithm from the SHAP Python package (LUNDBERG; LEE, 2017) in our RF classifier for the augmented datasets to compute the feature importance for each class and the impact on the model outputs. Figure 2 represents the workflow for the whole sound classification process, including the preprocessing, feature extraction and data augmentation steps (MOLNAR, 2019; LUNDBERG; LEE, 2017; SUNDARARAJAN; NAJMI, 2020) (For more details of the data augmentation process see electronic supplementary material, ‘Methods’.)

FIGURA 2 - Workflow for the feature selection and classification processes for reef fish sounds.



3. RESULTS

We have evaluated the performance of the models to determine their effectiveness in handling different feature representations, as well as the effects of data augmentation. Four metrics were used to evaluate the classification models (accuracy, precision, recall and f1-score), as shown in table 1. We observed that all classifiers had a satisfactory performance for all feature sets. In general, we observed the best results for RF and MLP, followed by NB for all feature sets. DT showed a less satisfactory performance; however, it performed well for feature set 1, with an accuracy of 91.7% and precision, recall and f1-scores of 92.0%.

TABELA 1 - Performance of supervised classification for the main original and augmented feature sets. The performance metrics are Ac (% accuracy), Pc (% precision), Rc (% recall) and f1 (% f1-score). Values in bold indicate the best performances.

Classifier	**Feature set**	**Original**				**Augmented**			
		Ac	**Pc**	**Rc**	**f1**	**Ac**	**Pc**	**Rc**	**f1**

NB	1	80.6	89.1	82.3	82.2	**93.5**	**93.5**	93.4	93.2
	2	80.6	77.8	77.4	76.7	**85.2**	85.1	84.4	84.5
	3	91.7	91.7	92.0	90.5	**95.4**	95.1	95.0	95.1
RF	1	88.9	90.1	89.9	90.0	**93.5**	93.1	93.0	93.0
	2	77.8	73.9	74.6	74.0	**88.9**	88.8	88.3	88.3
	3	83.3	86.4	86.6	84.1	**92.6**	92.4	91.9	92.1
DT	1	91.7	92.0	92.0	92.0	**92.6**	92.0	**92.6**	92.2
	2	63.3	63.5	63.0	62.2	**78.7**	77.7	77.7	77.6
	3	72.2	70.6	73.7	70.6	**88.9**	88.7	88.2	88.3
MLP	1	88.9	91.5	88.0	89.0	95.4	95.4	**95.6**	95.5
	2	75.0	78.6	75.2	74.3	90.7	**90.8**	90.5	90.4
	3	94.4	94.2	93.3	93.6	98.1	98.1	**98.2**	98.1

The data augmentation had a significant positive impact on the classifiers, as it increased the performance metrics. For instance, while the accuracy of the original feature set 3 was 94.4%, 98.1% of the calls were correctly classified by MLP after augmentation. The NB, MLP and RF classifiers were the most sensitive to data augmentation, while DT had less influence on its effects. Classification accuracy increased from 80.0% to 93.3% and from 88.9% to 93.5% for NB and RF from feature set 1, respectively. The cross-validation used to test our classifiers shows that the proposed architectures are satisfactory, with, for example, 95% ± 0.6% of correct classification by MLP (electronic supplementary material, table S3).

Regarding the feature sets, the third one was the most influential for all classifiers, reaching 98.1%accuracy for MLP, after augmentation. This set was the largest, with 40 features, and was the optimal feature set. Feature set 1 was the second most influential, with 95.4% of accuracy for the same classifier, after augmentation. The lowest performance for all classifiers was for feature set 2 compared with the other sets, but the correct classifications were superior to 75% for RF and MLP. It is worth mentioning that this set had a significant increase after augmentation, reaching 85.2%, 88.9%, 78.7% and 90.7% accuracy for NB, RF, DT and MLP, respectively. For the other features tested (electronic supplementary material, table S2), we observed that the TTFs set showed the best performance. The feature set with the second best performance was the MFCCs, followed by the spectral set. Temporal features presented the lowest performance metrics for the classifiers, but they increased significantly after augmentation.

We computed the SHAP values and the influence of each feature in predicting each class using the RF classifier for each set. The performance metrics (accuracy, precision, recall and f1-score) were also computed for each class for each feature set (electronic supplementary material, table S4). Accuracy was above 92% for all classes in sets 1 and 3, and above 86% for set 2, with Class 1 showing the highest accuracy in each set. For feature set 3—the optimal set—the most important features were interpulse interval (IPI), high-frequency, MFCC5, call duration and pulse duration (figure 3). In this feature set, we can see that high frequency and MFCC5 were the most important features for Class 3. IPI and call duration had a large contribution for Class 2. IPI and high frequency also contributed in predicting Classes 1 and 4.

FIGURA 3 - Summary of the contributions of features for each sound class for feature set 3.

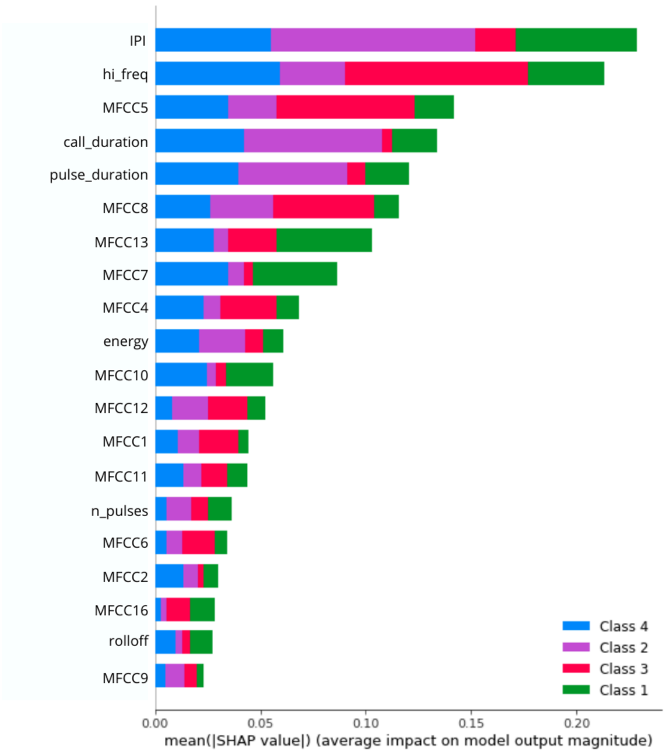
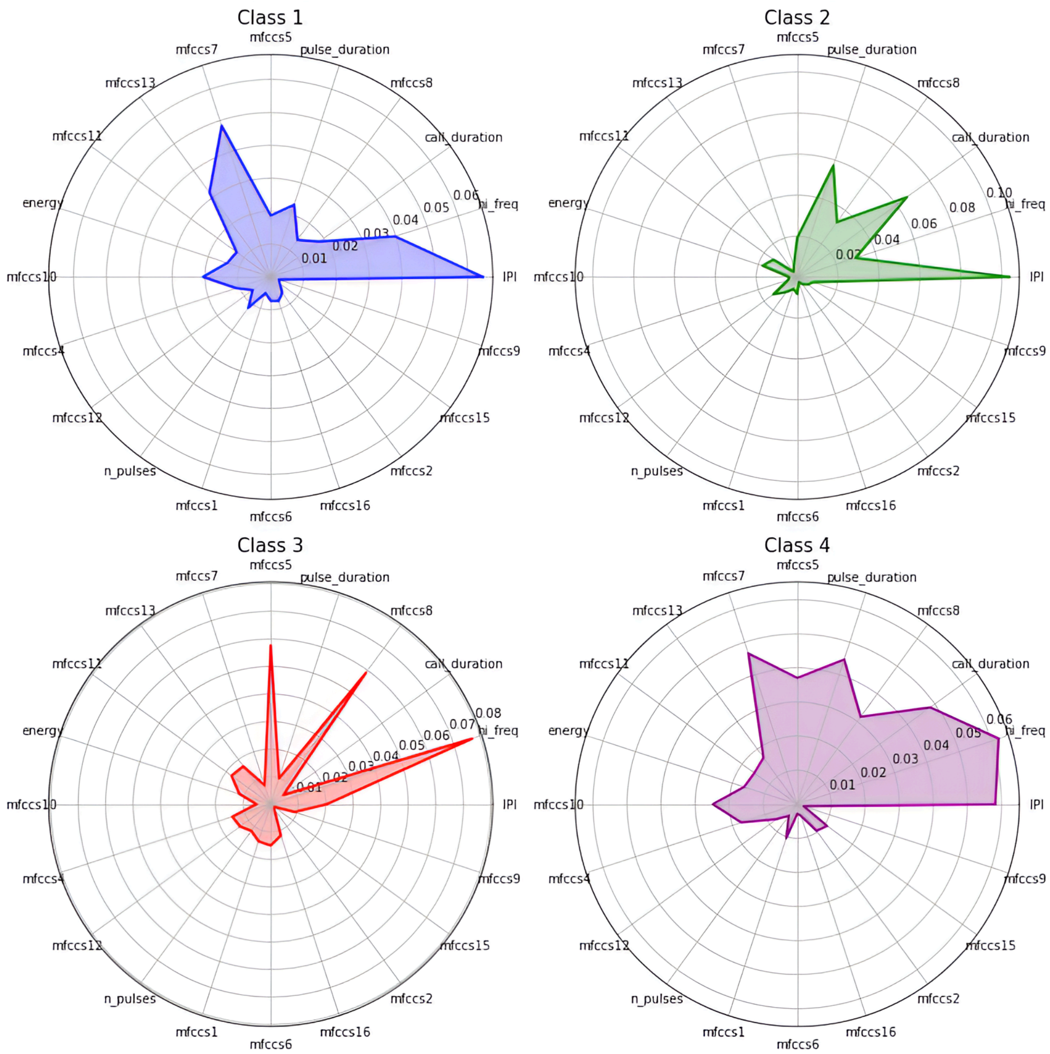


Figure 4 shows the polar plots for the SHAP values for each class for feature set 3, highlighting the features with the most significant contributions to the classification. For Class 1, IPI, MFCC7 and MFCC10 emerged as the key drivers, indicating the strong influence of these features in predicting this class. For Class 2, IPI, call duration and pulse duration were the most important features in distinguishing this class. High frequency, MFCC5 and MFCC8 were most influential in predicting Class 3, whereas IPI, high frequency, MFCC7 and pulse duration were most influential in predicting Class 4. These results highlight the different roles that specific features play in the predictions made by the model. Electronic supplementary material, figure S1 provides a different type of visualization for the summary of SHAP values for feature set 3, showing how the low or high values of the features impacted the outputs of the model. SHAP summary plots for each class for features set 1 and 2 are in the electronic supplementary material, figures S2 and S3, respectively.

FIGURA 4 - Polar plot for the summary of the SHAP values for each class for feature set 3. The amplitude of the polygons indicates the importance of features.



For TTFs, we used a dataset with MFCCs, spectral centroid, spectral rolloff, zero-crossing rate (ZCR) and spectral flux. The accuracy for this set was 85.2%, 88.9%, 77.8% and 91.7% after augmentation for NB, RF, DT and MLP, respectively (electronic supplementary material, table S2). This feature set showed the best performance among the other features tested. The main contributors to this performance were the MFCCs 8, 7, 5, as well as spectral centroid, ZCR and spectral rolloff. Spectral centroid influenced the prediction of Class 2, and ZCR and spectral rolloff contributed to the prediction of Class 1 (electronic supplementary material, figure S4). Spectral flux was not relevant for this set. However, for the set combining all of the spectral features, spectral flux was more important than centre and peak frequency (electronic supplementary material, figure S5).

4. DISCUSSION AND CONCLUSIONS

The results presented in this work show the effectiveness of the model in classifying unknown reef fish sounds across the feature spaces considered. The proposed classifiers were able to discriminate classes with similar temporal patterns (actually the same call type: pulse trains). Previous work has trained supervised algorithms to detect, identify or classify fish sounds using different algorithms (such as k-nearest neighbours, random forest and support vector machine), architectures and feature sets (NODA et al., 2016; MALFANTE et al., 2018; SATTAR et al., 2016; IBRAHIM et al., 2018). Their best performances were, respectively, 96.0%, 95.6%, 96.9% and 82.7%. Compared with the performance reported in these works, our results are promising, as the proposed models achieved similar performance. It is worth noting that we proposed a simpler approach with very small datasets and achieved excellent results, especially after augmentation.

MLP was the best classifier for these sounds, reaching an accuracy of 98.1% after augmentation for the optimal feature set. This classifier also showed high precision, recall and f1-score, which means a low number of false positives and false negatives (SALMAN et al., 2019). This is a robust algorithm, capable of estimating nonlinear functions and with the ability to learn hierarchical representations of data (REIFMAN; FELDMAN, 2022; BENGIO, 2022). RF was also a robust classifier, showing high accuracy for feature sets 1 and 3, especially after augmentation. This is an ensemble classifier with high accuracy and statistically robust results and it is currently one of the most powerful supervised algorithms (PARMAR et al., 2019; GÉRON, 2022). MLP and RF classifiers showed similar performance for the original datasets. The high performance of RF can be explained by the ability of tree-based classifiers to handle tabular data. In fact, these models are known to outperform deep learning for this type of data (GRINSZTAJN et al., 2022). However, after augmentation, MLP outperformed RF, as this technique optimizes the performance and robustness of deep learning models (ONISHI; MEGURO, 2023). On the other hand, DT was not a good classifier for this type of fish sound, as it had a lower performance for the optimal feature set and was the least sensitive to data augmentation. This classifier has some limitations such as instability, where changes in any variable or in the size of the training data can affect the classification rules (LI; BELFORD, 2002; KUMAR; KUMAR, 2020). Although NB showed higher performance, this algorithm may not be appropriate for this problem given the complexity of the fish sounds. Although it learns fast, it is a simpler algorithm, based on probability, which considers all variables as independent (DIMITOGLOU et al., 2012). Thus, the results show the importance of testing different classifiers, according to the specificities of the datasets.

Our results also demonstrated that data augmentation had an expressive influence on the classifiers, improving their performance. Data augmentation has been widely used in AI, especially in deep learning, which requires large datasets for training. This method is more commonly applied directly on images or raw audio files to significantly increase the size of datasets (STOWELL et al., 2019; STOWELL, 2022). Ibrahim et al. (2020) have used data augmentation directly on images of spectrograms to increase their dataset of grouper species. Waddell et al. (2021) also augmented their dataset of images of six different fish calls to optimize their classifier. Due to the complex structure of tabular data, data augmentation for this type of dataset is not well established and must be domain-specific (ONISHI; MEGURO, 2023; MACHADO et al., 2022). However, our approach using a comparatively small increase in the number of samples of tabular data (33%) could significantly improve the model performance. It is worth noting that in this work we used a balanced dataset (30 samples per class) to ensure that the classifiers were not biased towards a particular class. In addition, we performed cross-validation on both the original and augmented datasets, so the models could handle unseen data. The increase in performance demonstrated that this type of augmentation can be used in situations where large datasets of images or sound files are not available.

Feature set 3, with 40 features, was the optimal feature set with the best results (94.4% accuracy for the original and 98.1% for the augmented dataset with the MLP classifier). This is consistent with the results of Malfante et al. (2018) where the best results (96.9% and 96.5% for the random forest and support vector machine classifiers, respectively) were for their set with the highest number of features (84 features). In terms of the effect of data augmentation, feature set 2 was the most sensitive as it had the lowest values for the performance metrics, but after augmentation, the performance of the models for this set increased significantly. This can be explained by the small size of this set to train comparatively more complex features, thus requiring more data samples to train the classifiers. Regarding the other feature sets tested (electronic supplementary material, table S2), the temporal feature set showed the lowest performance. Despite being the same call type (pulse trains), it is worth mentioning that classes 1, 3 and 4 varied in the number of pulses within the classes, and not only between classes, which increased the complexity of the classification. However, performance increased after augmentation, as the models were able to learn with more examples.

The results of the feature contributions for each class for feature set 3 using the SHAP Explainer showed that Class 3 was most influenced by high frequency, as it had the lowest frequency of the classes tested (figure 3). For this class, high frequency acted as a 'threshold' as it had the lowest frequency of all classes. As high frequency and MFCC5 influenced the prediction for Class 3, we generated a dependence plot for this class (electronic supplementary material, figure S6) to investigate possible interaction effects of these two variables. The plot shows the influence of high frequency on the vertical distribution of SHAP values for MFCC5. Higher values of high frequency have a small relationship with MFCC5, indicating that this class, characterized by a low frequency band, is related to this coefficient. In general, MFCCs were important in predicting Class 3, which may be related to the distinct characteristics of the spectral envelope in the frequency band of this class. IPI and call duration influenced the predictions of Class 2. In fact, this class was the only one without within-class variation, with a very clear temporal pattern (always four pulses), making IPI and call duration good descriptors for this class. The dependence plot for this class for IPI and call duration (electronic supplementary material, figure S7) shows that higher values of call duration are associated with higher values of IPI.

For feature set 1, high frequency, IPI and call duration were the most important features. High frequency had more influence in predicting Classes 3 and 4, while IPI and call duration had more influence in predicting Class 2 (electronic supplementary material, figure S8). In feature set 2, we can observe that the most important features were the MFCCs 7, 5 and 8. Although the temporal feature set had a lower performance, IPI, pulse duration and call duration showed significant contribution in predicting Class 1. For feature set 3, the summary plot for Class 1 (electronic supplementary material, figure S1a) shows that higher values of these features were one of the most important features in predicting this class. In general, the MFCCs had more influence in predicting Class 3 (electronic supplementary material, figure S1c), followed by Classes 4 and 1, and less influence on the prediction of Class 2 (electronic supplementary material, figure S9). Regarding the spectral feature set, it is possible to observe that high frequency and TTFs (spectral rolloff, spectral centroid and zero crossing rate) were the most important in predicting classes (electronic supplementary material, figure S5). The most important MFCCs by class were MFCC5, MFCC8 and MFCC4 for Class 3 (100−400 Hz); MFCC8 for Class 2 (100–600 Hz); MFCC13 for Class 1 (100–500 Hz); and MFCC7 for Classes 1 and 4 (100–500 and 100−700 Hz, respectively). Higher-order MFCCs (i.e. >14) were not significant in discriminating between classes.

Owing to the effectiveness of MFCCs in lower frequencies, these features are suitable for most fish sounds. This explains the recent popularity of MFCCs in the field of fish sounds. Previous studies have included MFCCs in their feature sets to detect, classify and identify fish sounds (RUIZ-BLAIS et al., 2012; VIEIRA et al., 2015; NODA et al., 2016; SATTAR et al., 2016; MALFANTE et al., 2018; HARAKAWA et al., 2018; IBRAHIM et al., 2018; VIEIRA et al., 2019), achieving satisfactory results for their models. However, they have not provided a detailed discussion on the contribution of each coefficient for specific sound classes. The individual analysis of each coefficient may offer a more refined spectral representation of the signals. TTFs have also been explored in previous work for fish sound detection and classification (RUIZ-BLAIS et al., 2012; VIEIRA et al., 2015; HARAKAWA et al., 2018). For example, Ruiz-Blais et al. (2012) used statistics of different timbral features to detect sciaenid species, achieving 89% correctness for their set with all features tested, after post-processing. In their work, they found that cepstral coefficients, spectral centroid and zero-crossing rate were the most relevant. Our results are consistent with theirs because the same features were most relevant, but we achieved 91.7% of accuracy for the augmented set for the MLP classifier. These results demonstrated that TTFs were also good descriptors for these sound types (see electronic supplementary material, table S2 and figure S4).

Fish are known to use sound for many purposes, including mating, territorial defence, feeding, schooling behaviour and many others (AMORIM, 2006; LADICH, 2019). However, a current limitation of fish sound classification is the lack of data to compare and match unknown sounds to specific species. Although over 1000 species from 175 families (including Haemulidae, Holocentridae, Batrachoididae, Serranidae, Sciaenidae, Pomacanthidae, Pomacentridae, among many others) are known to produce sounds (LOBEL et al., 2010; RICE, 2022; LOOBY, 2023), the repertoire of most soniferous species has not been totally characterized (PARSONS et al., 2022). This gap in the acoustic ecology of fish is still a challenge for an ecological application of PAM in terms of identification and characterization of fish biodiversity patterns and to understand aspects of species behaviour. Although it was not possible to identify the species producing the sounds used in this paper, our results demonstrate the effectiveness of artificial intelligence in classifying pulsed fish sounds. Classes 2 and 3 show similarities to sounds described for the Pomacentridae family, especially those produced by the genus Stegastes (SPANIER, 1979; FRÉDÉRICH; PARMENTIER, 2016). In summary, for those classes of sounds possibly produced by Stegastes, the best features that described these sounds were IPI, call duration and pulse duration (Class 2) and high frequency, MFCC5 and MFCC8 (Class 3). Given that some species in this genus are known for their territorial behaviour and the crucial role that sound plays in agonistic interactions related to territory and nest defence in coral reefs (FRÉDÉRICH; PARMENTIER, 2016), advances in the detection of these sounds are crucial. Improved sound classification can provide valuable insights into the dynamics of this keystone group and enhance our understanding of their ecological roles and interactions within their habitats.

Explainable AI allowed us to identify which features contributed to each sound class. Introducing the interpretability of the models in the complex problem of fish sound classification is an important step, as it provides insights into which features shape a particular sound type. In addition, it is important to understand the mechanisms that lead the model to achieve certain performance metrics in order to assess its reliability (RIBEIRO et al., 2016). XAI also enabled us to evaluate the impact of each MFCC in predicting each class. MFCCs are widely used in speech recognition to study emotions—and even diseases (such as depression)—with distinct coefficients used as indicators of specific disorders (WANG et al., 2019; TRACEY et al., 2023). For fish, these may be important descriptors for recognizing species-specific sounds or indicators of behaviours. Recognizing and characterizing these sounds are key to better understanding diel behaviours and functional roles linked to critical reef ecological processes (BELLWOOD et al., 2019). This is a significant advance because most of the studies using AI to detect and classify fish sounds have focused only on target species and the majority of these sounds remain unidentified and uncharacterized in current research (PARSONS et al., 2022; MOUY et al., 2024). There is therefore a growing need to develop reliable detectors and classifiers that are capable of interpreting unknown fish sounds. By focusing on the development of such advanced detectors and classifiers, we can improve our ability to categorize and understand a wider range of fish acoustic signals. The ultimate knowledge of the spatial and temporal patterns of biophony and associated underwater soundscape are the next generation of data for the optimal management and conservation of marine ecosystems and associated biodiversity (DUARTE et al., 2021).

5. REFERENCES

ABAYOMI-ALLI, O. O.; DAMAŠEVIČIUS, R.; QAZI, A.; ADEDOYIN-OLOWE, M.; MISRA, S. Data augmentation and deep learning methods in sound classification: a systematic review. *Electronics*, v. 11, p. 3795, 2022.

AMORIM, M. Diversity of sound production in fish. In: LADICH, F.; COLLIN, S. P.; MOLLER, P.; KAPOOR, B. G. (ed.). *Communication in fishes*. Enfield: Science Publishers, 2006. p. 71–104.

BARROSO, V. R.; XAVIER, F. C.; FERREIRA, C. E. L. Applications of machine learning to identify and characterize the sounds produced by fish. *ICES Journal of Marine Science*, v. 80, p. 1854–1867, 2023.

BARROSO, V. R.; LESSA, A.; FERREIRA, C. E. L.; XAVIER, F. Supplementary material from: Optimal feature selection and model explanation for reef fish sound classification. *Figshare*, 2025. Disponível em: https://doi.org/10.6084/m9.figshare.c.7837885.

BELLWOOD, D. R.; STREIT, R. P.; BRANDL, S. J.; TEBBETT, S. B. The meaning of the term “function” in ecology: a coral reef perspective. *Functional Ecology*, v. 33, p. 948–961, 2019.

BENGIO, Y. Deep learners benefit more from out-of-distribution examples. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 14., 2011, Fort Lauderdale. *Proceedings…* Fort Lauderdale: JMLR Workshop and Conference Proceedings, 2011. p. 164–172.

BIANCO, M. J.; GERSTOFT, P.; TRAER, J.; OZANICH, E.; ROCH, M. A.; GANNOT, S.; DELEDALLE, C. A. Machine learning in acoustics: theory and applications. *The Journal of the Acoustical Society of America*, v. 146, p. 3590–3628, 2019.

BOLGAN, M.; CRUCIANELLI, A.; MYLONAS, C. C.; HENRY, S.; FALGUIÈRE, J. C.; PARMENTIER, E. Calling activity and calls’ temporal features inform about fish reproductive condition and spawning in three cultured Sciaenidae species. *Aquaculture*, v. 524, p. 735243, 2020.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Regression trees. In: BREIMAN, L. et al. *Classification and regression trees*. 1. ed. New York: Chapman and Hall/CRC, 2017. p. 216–265.

CAI, W.; ZHU, J.; ZHANG, M.; YANG, Y. A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation. *Sensors*, v. 22, p. 7443, 2022.

CAMPBELL, D.; XAVIER, F. C.; MELO JUNIOR, U. G.; SILVEIRA, N. G.; VERSIANI, L. L.; NETTO, E. B. F. Underwater soundscape pattern during high season of nautical tourism in Cabo Frio Island, Brazil. In: INTERNATIONAL CONFERENCE ON THE EFFECTS OF NOISE ON AQUATIC LIFE, 5., 2019, Den Haag. *Proceedings of Meetings on Acoustics*. Melville: AIP Publishing, v. 37, p. 070003, 2019.

CARRIÇO, R.; SILVA, M. A.; MENESES, G. M.; FONSECA, P. J.; AMORIM, M. C. P. Characterization of the acoustic community of vocal fishes in the Azores. *PeerJ*, v. 7, p. e7772, 2019.

CARRIERE, O.; HERMAND, J. P.; CALADO, L.; DE PAULA, A. C.; DA SILVEIRA, I. C. A. Feature-oriented acoustic tomography: upwelling at Cabo Frio (Brazil). In: OCEANS 2009 MTS/IEEE BILOXI – MARINE TECHNOLOGY FOR OUR FUTURE: GLOBAL AND LOCAL CHALLENGES, 2009, Biloxi. *Proceedings…* Biloxi: IEEE, 2009. p. 1–8.

CHÉRUBIN, L. M.; DALGLEISH, F.; IBRAHIM, A. K.; SCHÄRER-UMPIERRE, M.; NEMETH, R. S.; MATTHEWS, A.; APPELDOORN, R. Fish spawning aggregations dynamics as inferred from a novel, persistent presence robotic approach. *Frontiers in Marine Science*, v. 6, p. 779, 2020.

CHARIF, R. A.; STRICKMAN, L. M.; WAACK, A. M. *Raven Pro 1.4 user's manual*. Ithaca: The Cornell Lab of Ornithology, 2010. Disponível em: https://www.ravensoundsoftware.com/knowledge-base/.

CORDEIRO, C. A. M. M.; HARBORNE, A. R.; FERREIRA, C. E. L. Patterns of distribution and composition of sea urchin assemblages on Brazilian subtropical rocky reefs. *Marine Biology*, v. 161, p. 2221–2232, 2014.

CORDEIRO, C. A. M. M.; MENDES, T. C.; HARBORNE, A. R.; FERREIRA, C. E. L. Spatial distribution of nominally herbivorous fishes across environmental gradients on Brazilian rocky reefs. *Journal of Fish Biology*, v. 89, p. 939–958, 2016.

DESIDERÀ, E.; GUIDETTI, P.; PANZALIS, P.; NAVONE, A.; VALENTINI-POIRRIER, C.; BOISSERY, P.; GERVAISE, C.; DI IORIO, L. Acoustic fish communities: sound diversity of rocky habitats reflects fish species diversity. *Marine Ecology Progress Series*, v. 608, p. 183–197, 2019.

DIMITOGLOU, G.; ADAMS, J. A.; JIM, C. M. Comparison of the C4.5 and a naïve Bayes classifier for the prediction of lung cancer survivability. *arXiv*, 2012. Disponível em: https://arxiv.org/abs/1206.1121.

DUARTE, C. M. et al. The soundscape of the Anthropocene ocean. *Science*, v. 371, p. eaba4658, 2021.

FERREIRA, C.; GONÇALVES, J.; COUTINHO, R. Community structure of fishes and habitat complexity on a tropical rocky shore. *Environmental Biology of Fishes*, v. 61, p. 353–369, 2001.

FISH, M. P.; MOWBRAY, W. H. *Sounds of western North Atlantic fishes: a reference file of biological underwater sounds*. Baltimore: Johns Hopkins Press, 1970.

FRÉDÉRICH, B.; PARMENTIER, E. *Biology of damselfishes*. 1. ed. Boca Raton: CRC Press, 2016.

FROESE, R.; PAULY, D. *FishBase*. 2021. Disponível em: https://www.fishbase.org.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. 3. ed. Sebastopol: O'Reilly Media, 2022.

GIANNAKOPOULOS, T.; PIKRAKIS, A. *Introduction to audio analysis: a MATLAB® approach*. Oxford: Academic Press, 2014.

GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, v. 35, p. 507–520, 2022.

HARAKAWA, R.; OGAWA, T.; HASEYAMA, M.; AKAMATSU, T. Automatic detection of fish sounds based on multi-stage classification including logistic regression via adaptive feature weighting. *The Journal of the Acoustical Society of America*, v. 144, p. 2709–2718, 2018.

HULSEN, T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. *AI*, v. 4, p. 652–666, 2023.

IBRAHIM, A. K.; CHÉRUBIN, L. M.; ZHUANG, H.; SCHÄRER-UMPIERRE, M. T.; DALGLEISH, F.; ERDOL, N.; OUYANG, B.; DALGLEISH, A. An approach for automatic classification of grouper vocalizations with passive acoustic monitoring. *The Journal of the Acoustical Society of America*, v. 143, p. 666–676, 2018.

IBRAHIM, A. K.; ZHUANG, H.; CHÉRUBIN, L. M.; SCHÄRER-UMPIERRE, M. T.; NEMETH, R. S.; ERDOL, N.; ALI, A. M. Transfer learning for efficient classification of grouper sound. *The Journal of the Acoustical Society of America*, v. 148, p. EL260–EL266, 2020.

ISHIKAWA, S.-N.; TODO, M.; TAKI, M.; UCHIYAMA, Y.; MATSUNAGA, K.; LIN, P.; OGIHARA, T.; YASUI, M. Example-based explainable AI and its application for remote sensing image classification. *International Journal of Applied Earth Observation and Geoinformation*, v. 118, p. 103215, 2023.

JASIM, W. N.; SADDAM, S. A. W.; HARFASH, E. J. Wind sounds classification using different audio feature extraction techniques. *Informatica*, v. 45, p. 57–65, 2022.

JESUS, S. M.; XAVIER, F. C.; VIO, R. P.; OSOWSKY, J.; SIMÕES, M. V. S.; NETTO, E. B. F. Particle motion measurements near a rocky shore off Cabo Frio Island. *The Journal of the Acoustical Society of America*, v. 147, p. 4009–4019, 2020.

KASUMYAN, A. O. Acoustic signaling in fish. *Journal of Ichthyology*, v. 49, p. 963–1020, 2009.

KUMAR, P.; KUMAR, D. Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, v. 12, p. 246–269, 2020.

LADICH, F. Fish bioacoustics. *Current Opinion in Neurobiology*, v. 28, p. 121–127, 2014.

LADICH, F. Ecology of sound communication in fishes. *Fish and Fisheries*, v. 20, p. 552–563, 2019.

LAPLANTE, J. F.; AKHLOUFI, M. A.; GERVAISE, C. Deep learning for marine bioacoustics and fish classification using underwater sounds. In: IEEE CANADIAN CONFERENCE ON ELECTRICAL AND COMPUTER ENGINEERING (CCECE), 2022, Halifax. *Proceedings…* Halifax: IEEE, 2022. p. 288–293.

LEWIS, D. D. Naive (Bayes) at forty: the independence assumption in information retrieval. In: NÉDELLEC, C.; ROUVEIRO, C. (ed.). *Machine learning: ECML-98*. Berlin: Springer, 1998. p. 4–15. (Lecture Notes in Computer Science, v. 1398).

LI, R. H.; BELFORD, G. G. Instability of decision tree classification algorithms. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 8., 2002, Edmonton. *Proceedings…* New York: Association for Computing Machinery, 2002. p. 570–575.

LIMA, J. A. M. et al. The Brazilian Santos basin underwater soundscape monitoring project (PMPAS-BS). *Frontiers in Marine Science*, v. 11, p. 1416590, 2024.

LISA YANG CENTER FOR CONSERVATION BIOACOUSTICS AT THE CORNELL LAB OF ORNITHOLOGY. *Raven Pro: interactive sound analysis software*. Ithaca: The Cornell Lab of Ornithology, 2019. Disponível em: https://www.ravensoundsoftware.com/software/raven-pro/.

LOBEL, P. S.; KAATZ, I. M.; RICE, A. N. Acoustical behavior of coral reef fishes. In: COLE, K. (ed.). *Reproduction and sexuality in marine fishes: patterns and processes*. Oakland: California Scholarship Online, 2010. p. 307–386.

LOOBY, A. FishSounds version 1.0: a website for the compilation of fish sound production information and recordings. *Ecological Informatics*, v. 74, p. 101953, 2023.

LUNDBERG, S. M.; LEE, S. I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, v. 30, p. 4768–4777, 2017.

MACHADO, P.; FERNANDES, B.; NOVAIS, P. Benchmarking data augmentation techniques for tabular data. In: YIN, H.; CAMACHO, D.; TINO, P. (ed.). *International Conference on Intelligent Data Engineering and Automated Learning*, 2022, Manchester. *Proceedings…* Cham: Springer International Publishing, 2022. p. 104–112.

MALFANTE, M.; MARS, J. I.; DALLA MURA, M.; GERVAISE, C. Automatic fish sounds classification. *The Journal of the Acoustical Society of America*, v. 143, p. 2834–2846, 2018.

MANN, D.; GROTHUES, T. Short-term upwelling events modulate fish sound production at a mid-Atlantic Ocean observatory. *Marine Ecology Progress Series*, v. 375, p. 65–71, 2009.

MCFEE, B.; COLINR, R.; DAWEN, L.; ELLIS, D. P.; MCVICAR, M.; BATTENBERG, E.; NIETO, E. *librosa: audio and music signal analysis in Python*. In: HUFF, K.; BERGSTRA, J. (ed.). PYTHON IN SCIENCE CONFERENCE (SCIPY), 14., 2015, Austin. *Proceedings…* Austin: SciPy, 2015. p. 18–24.

MCWILLIAM, J. N.; MCCAULEY, R. D.; ERBE, C.; PARSONS, M. J. G. Patterns of biophonic periodicity on coral reefs in the Great Barrier Reef. *Scientific Reports*, v. 7, p. 17459, 2017.

MOLNAR, C. *Interpretable machine learning: a guide for making black box models explainable*. 2. ed. 2019. Disponível em: https://christophm.github.io/interpretable-ml-book/.

MOUY, X. et al. Automatic detection of unidentified fish sounds: a comparison of traditional machine learning with deep learning. *Frontiers in Remote Sensing*, v. 22, p. 1439995, 2024.

NODA, J.; TRAVIESO, C.; SÁNCHEZ-RODRÍGUEZ, D. Automatic taxonomic classification of fish based on their acoustic signals. *Applied Sciences*, v. 6, p. 443, 2016.

ONISHI, S.; MEGURO, S. Rethinking data augmentation for tabular data in deep learning. *arXiv*, 2023. Disponível em: https://arxiv.org/abs/2305.10308.

PARCERISAS, C.; ROCA, I. T.; BOTTELDOOREN, D.; DEVOS, P.; DEBUSSCHERE, E. Categorizing shallow marine soundscapes using explained clusters. *Journal of Marine Science and Engineering*, v. 11, p. 550, 2023.

PARMAR, A.; KATARIYA, R.; PATEL, V. A review on random forest: an ensemble classifier. In: HEMANTH, J.; FERNANDO, X.; LAFATA, P.; BAIG, Z. (ed.). *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI)*, 2018. Cham: Springer International Publishing, 2019. p. 758–763. (Lecture Notes on Data Engineering and Communications Technologies, v. 26).

PARSONS, M. J. G. et al. Sounding the call for a global library of underwater biological sounds. *Frontiers in Ecology and Evolution*, v. 10, p. 39, 2022.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PUTLAND, R.; MACKIEWICZ, A.; MENSINGER, A. Localizing individual soniferous fish using passive acoustic monitoring. *Ecological Informatics*, v. 48, p. 60–68, 2018.

RAO, K.; MANJUNATH, K. *Speech recognition using articulatory and excitation source features*. Fort Lee: Springer, 2017.

REIFMAN, J.; FELDMAN, E. Multilayer perceptron for nonlinear programming. *Computers & Operations Research*, v. 29, p. 1237–1250, 2002.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. “Why should I trust you?” Explaining the predictions of any classifier. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. *Proceedings…* New York: Association for Computing Machinery, 2016. p. 1135–1144.

RICE, A. N. Evolutionary patterns in sound production across fishes. *Ichthyology & Herpetology*, v. 110, p. 1–12, 2022.

RICE, A. N.; SOLDEVILLA, M. S.; QUINLAN, J. A. Nocturnal patterns in fish chorusing off the coasts of Georgia and eastern Florida. *Bulletin of Marine Science*, v. 93, p. 455–474, 2017.

RUIZ-BLAIS, S.; RIVERA-CHAVARRIA, M. R.; CAMACHO, A. Autonomous detection of neotropical sciaenid fishes. *Proceedings of Meetings on Acoustics*, v. 18, p. 010001, 2012.

RUMELHART, D. E.; MCCLELLAND, J. L. A general framework for parallel distributed processing. In: RUMELHART, D. E.; MCCLELLAND, J. L. (ed.). *Parallel distributed processing: explorations in the microstructure of cognition*. Cambridge: MIT Press, 1987. p. 45–76.

SALMAN, A.; MAQBOOL, S.; KHAN, A.; JALAL, A.; SHAFAIT, F. Real-time fish detection in complex backgrounds using probabilistic background modelling. *Ecological Informatics*, v. 51, p. 44–51, 2019.

SATTAR, F.; CULLIS-SUZUKI, S.; JIN, F. Identification of fish vocalizations from ocean acoustic data. *Applied Acoustics*, v. 110, p. 248–255, 2016.

SHARMA, G.; UMAPATHY, K.; KRISHNAN, S. Trends in audio signal feature extraction methods. *Applied Acoustics*, v. 158, p. 107020, 2020.

SIDDAGANGAIAH, S.; CHEN, C. F.; HU, W. C.; DANOVARO, R.; PIERETTI, N. Silent winters and rock-and-roll summers: the long-term effects of changing oceans on marine fish vocalization. *Ecological Indicators*, v. 125, p. 107456, 2021.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, p. 427–437, 2009.

SPANIER, E. Aspects of species recognition by sound in four species of damselfishes, genus *Eupomacentrus* (Pisces: Pomacentridae). *Zeitschrift für Tierpsychologie*, v. 51, p. 301–316, 1979.

STOWELL, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, v. 10, p. e13152, 2022.

STOWELL, D.; PETRUSKOVÁ, T.; ŠÁLEK, M.; LINHART, P. Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *Journal of the Royal Society Interface*, v. 16, p. 20180940, 2019.

SUNDARARAJAN, M.; NAJMI, A. The many Shapley values for model explanation. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML), 37., 2020, Online. *Proceedings of Machine Learning Research*, v. 119, p. 9269–9278, 2020. Disponível em: https://proceedings.mlr.press/v119/sundararajan20b.html.

TORRES-GARCÍA, A.; GARCIA, C.; VILLASEÑOR-PINEDA, L.; MENDOZA-MONTOYA, O. (ed.). *Biosignal processing and classification using computational learning and intelligence: principles, algorithms, and applications*. London: Academic Press, 2021.

TRACEY, B. et al. Towards interpretable speech biomarkers: exploring MFCCs. *Scientific Reports*, v. 13, p. 22787, 2023.

VALENTIN, J. L. Analyse des paramètres hydrobiologiques dans la remontée de Cabo Frio (Brésil). *Marine Biology*, v. 82, p. 259–276, 1984.

VIEIRA, M.; FONSECA, P. J.; AMORIM, M. C. P.; TEIXEIRA, C. J. C. 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*, v. 138, p. 3941–3950, 2015.

VIEIRA, M.; PEREIRA, B. P.; POUSÃO-FERREIRA, P.; FONSECA, P. J.; AMORIM, M. C. P. Seasonal variation of captive meagre acoustic signalling: a manual and automatic recognition approach. *Fishes*, v. 4, p. 28, 2019.

VIVIANEOCN. *fish_sounds_shap*. GitHub, 2024. Disponível em: https://github.com/vivianeocn/fish_sounds_shap.

WADDELL, E. E.; RASMUSSEN, J. H.; ŠIROVIĆ, A. Applying artificial intelligence methods to detect and classify fish calls from the Northern Gulf of Mexico. *Journal of Marine Science and Engineering*, v. 9, p. 1128, 2021.

WALL, C.; SIMARD, P.; LEMBKE, C.; MANN, D. Large-scale passive acoustic monitoring of fish sound production on the West Florida Shelf. *Marine Ecology Progress Series*, v. 484, p. 173–188, 2013.

WANG, J.; ZHANG, L.; LIU, T.; PAN, W.; HU, B.; ZHU, T. Acoustic differences between healthy and depressed people: a cross-situation study. *BMC Psychiatry*, v. 19, p. 1–12, 2019.

ZELKOWITZ, M. (ed.). *Advances in computers: improving the web*. v. 78. Oxford: Academic Press, 2010.

4. CAPÍTULO 3 - EXPLORING THE FEASIBILITY OF MULTIMODAL EMBEDDINGS FOR FISH SOUND CLASSIFICATION

ABSTRACT

Passive acoustic monitoring has gained increasing relevance in fish bioacoustics. However, manually analyzing large volumes of acoustic data is unfeasible. In addition, the application of traditional machine learning and deep learning models is further constrained by the scarcity of species-specific labeled datasets of fish sounds. This study investigates the use of zero-shot classification for fish acoustic signals using Contrastive Language–Audio Pretraining (CLAP), a multimodal architecture that maps audio and text into a shared semantic space. We assembled a labeled dataset from multiple sound libraries and conducted two experimental scenarios (structured vs descriptive prompts) to evaluate the performance of the model across different natural language query construction strategies. Cross-modal retrieval was employed to compute the similarity between audio and text embeddings, enabling the identification of the most likely acoustic category. The results highlight the potential of multimodal embeddings, even in the absence of domain-specific pretraining. For the proposed methodology, the findings highlight that prompt engineering influences the performance of the model, with descriptive prompts outperforming the structured ones in zero-shot classification. Notably, the performance improved with broader taxonomic levels, demonstrating a greater capacity to recognize acoustic patterns at the family level. The Sciaenidae, in particular, showed the strongest performance, accounting for over 60% of all correct classifications. Moreover, for this family, the model correctly retrieved at least one candidate in the Top-5 for all but one of the represented species. These findings suggest that zero-shot multimodal approaches can effectively support the development of more flexible and taxonomically informed AI tools for large scale marine acoustic monitoring, especially in data-scarce scenarios.

Keywords: fish sounds, zero-shot classification, multimodal embeddings, CLAP

1. INTRODUCTION

Over one thousand fish species are known to actively produce sounds (RICE et al., 2022; LOOBY et al., 2024) in various behavioral contexts, including reproduction, feeding, and agonistic interactions (KASUMYAN, 2009). In this context, Passive Acoustic Monitoring (PAM) has gained popularity in marine research and fish bioacoustics (LANCASTER et al., 2025), as long-term recordings can capture continuous biological activity and provide insights into fish behavior and relevant information to fisheries management, such as spawning patterns (GANNON, 2008). However, the environmental complexity of the ocean, marked by low signal-to-noise ratios, overlapping sounds (HILDEBRAND, 2009), and the large volume of data generated by PAM recordings, makes manual analysis impractical (BIANCO et al., 2019). Therefore, adopting robust computational methods is not only advantageous but essential for extracting biologically relevant information and enabling long-term monitoring of different marine ecosystems.

To address the challenges of data volume, various artificial intelligence (AI) techniques have been applied to detect and classify fish sounds (BARROSO et al., 2023). Initial efforts relied on machine learning with manually extracted features (BIANCO et al., 2019). These efforts progressed toward deep learning models, such as convolutional neural networks (CNNs), which automatically learn acoustic features (GOODWIN et al., 2022). More recently, Explainable AI (XAI) approaches have been explored to provide transparency for these models and to identify specific descriptors for fish sounds (BARROSO et al., 2025). While these methods have advanced the field, they remain heavily dependent on large amounts of species-specific training sets, which are rarely available for most marine species (PARSONS et al., 2022).

Recent advances in multimodal learning represent the current state-of-the-art (SotA) in AI and facilitate the representation of linguistic elements in vector spaces (ZHANG et al., 2024). Unlike unimodal systems, multimodal models such as Contrastive Language-Audio Pretraining (CLAP) integrate audio and text into a shared semantic space (WU et al., 2023). This architecture allows the model to use cross-modal embeddings, which are high-dimensional vector representations. In this study, these embeddings are used to map acoustic patterns and textual descriptions into the same space to measure their similarity. A key advantage of this architecture is its zero-shot learning capability, which is a technique that allows the model to categorize novel samples without the need for extensive, species-specific labeled datasets. By leveraging cross-modal representations learned during large-scale pretraining on various datasets (RADFORD et al., 2021), these models can generalize to unseen categories (ELIZALDE et al., 2023). This can help address the data scarcity gap in marine bioacoustics. Although these multimodal embeddings have been widely used in human speech and general audio recognition, their application in bioacoustics is still in its early stages (MIAO et al., 2025).

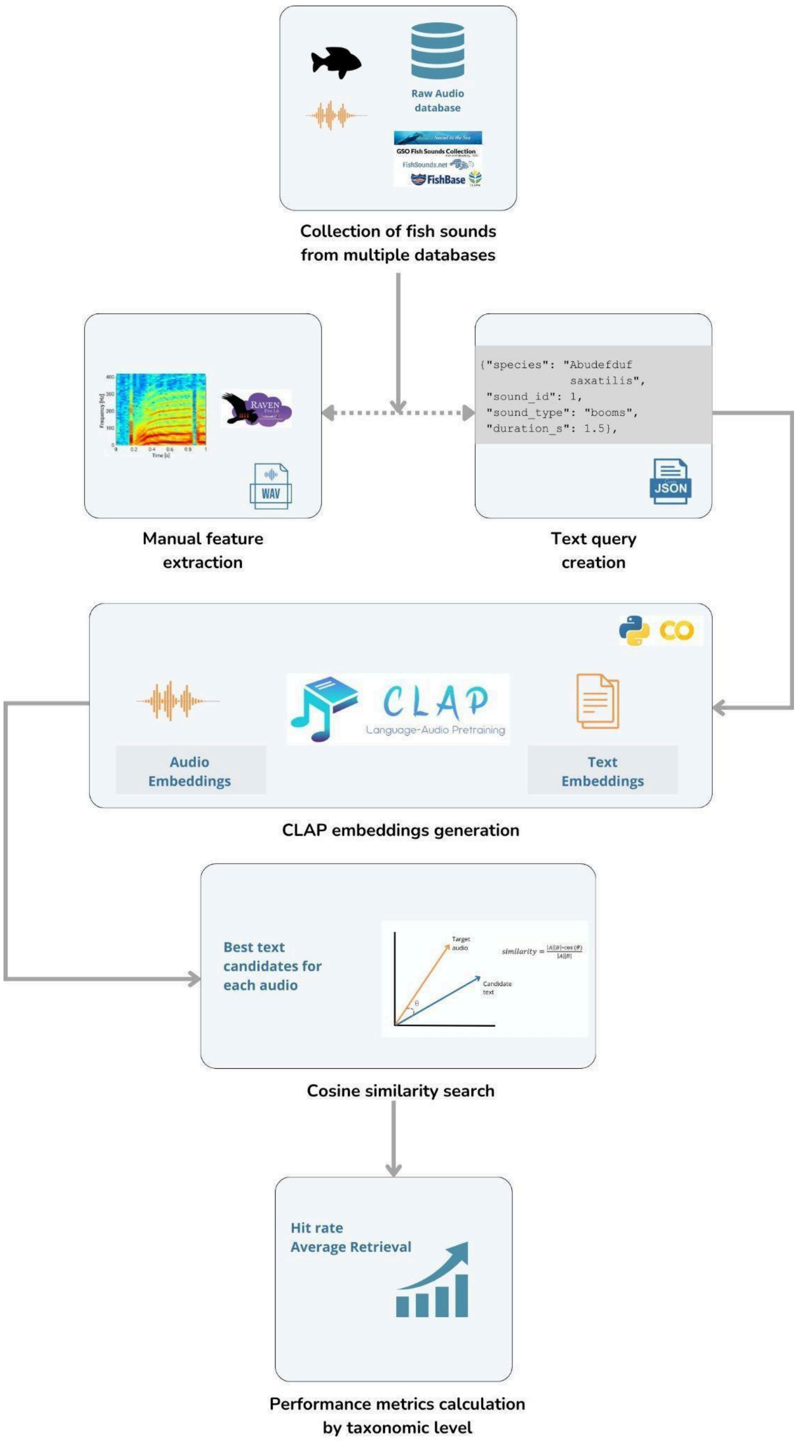
In this paper, we explore zero-shot classification of fish sounds using the CLAP model. This work is among the first to explore the use of multimodal embeddings to classify fish sounds, starting from a pre-labeled database. Although the overall performance is constrained by the fact that the model was not originally trained on bioacoustic data, the results nonetheless highlight the strong potential of the proposed framework. The present study aims to evaluate the performance of the model across different taxonomic levels using a dataset assembled from multiple global sound libraries to represent diverse marine and conducted two experimental scenarios to compare distinct natural language query construction strategies. Notably, the performance improved with broader taxonomic levels, demonstrating a greater capacity to recognize acoustic patterns at the family level. In particular, the model demonstrated a consistent ability to capture acoustic patterns within the Sciaenidae family (croakers and drums), a taxon of high ecological and economic relevance (SANTOS et al., 2012), that is also an acoustically relevant component of coastal soundscapes (MONCZAK et al., 2019). The successful identification of distinct acoustic patterns across different taxonomic levels suggests that, even without domain-specific pretraining, such approaches can support the development of more versatile and taxonomically informed AI tools for marine acoustic monitoring.

2. MATERIALS AND METHODS

The workflow developed in this study is structured into four main stages, as illustrated in Figure 1. First, a diverse acoustic dataset was assembled and curated from global sound libraries, followed by an exploratory data analysis (EDA) of the acoustic parameters of the fish signals. The core methodological framework consists of the zero-shot classification approach and the design of natural language queries.

Finally, the performance of the model was evaluated across different taxonomic levels.

FIGURA 1 - Workflow for fish sound classification using Contrastive Language-Audio Pretraining (CLAP). The process includes sound collection, feature extraction, text query creation, generation of audio and text embeddings, retrieval via cosine similarity and model evaluation through performance metrics by taxonomic levels.



2.1. Dataset compilation and curation

The dataset used in this paper was constructed by collecting sound examples from existing repositories, such as FishBase (FROESE; PAULY, 2024), Fish Sounds (LOOBY et al., 2024), Discovery of Sound in the Sea (DOSITS), and the Graduate School of Oceanography (GSO) Fish Sound Collection – University of Rhode Island (FISH; MOWBRAY, 1970). The dataset also includes fish sounds recorded by researchers from the Instituto de Estudos do Mar Almirante Paulo Moreira (IEAPM) (XAVIER, 2021).

The audio files were first converted to the *.wav* format, then edited using Audacity® software to remove human voices and other unwanted noises, while retaining only the signals produced by fishes. Following the audio editing process, spectrograms were generated using Raven Pro 1.6 (K. Lisa Yang Center for Conservation Bioacoustics, 2023). These spectrograms were then analyzed, and basic acoustic features were manually extracted (e.g., minimum and maximum frequencies and duration). Based on this acoustic dataset, a spreadsheet was created containing information on the annotated sound types produced by each species and the features extracted, as well as metadata (e.g., species name, common name, sampling rate, bit depth, and number of channels) for each audio file.

2.2. Exploratory Data analysis

An EDA of the constructed dataset was conducted using a Python code. This analysis aimed to extract information on the distribution of basic spectral and temporal acoustic parameters for each sound and to characterize the taxonomic groups (e.g., species, genus and family) present in the dataset.

Descriptive statistics were computed for numerical variables to summarize the measures of central tendency and dispersion, as well as to provide an overview of the variability present in the acoustic parameters. Categorical variables, such as fish family and source library, were examined by identifying unique categories and calculating their frequency distributions. Kernel density estimation (KDE) plots were produced for minimum and maximum frequencies and duration to allow for a smoother representation of the underlying probability density functions.

This exploratory analysis supported the interpretation of the multimodal classification results by highlighting acoustic variability and taxonomic patterns within the dataset, particularly in the context of zero-shot learning and cross-modal retrieval.

2.3. Zero-shot classification with multimodal embeddings

2.3.1. The CLAP framework

The CLAP model is a multimodal framework designed to learn audio representation by combining audio data and natural language descriptions through contrastive learning (WU et al., 2023). Instead of relying on conventional audio–label datasets, CLAP is trained on paired audio and textual descriptions. These audio–text pairs are drawn from a diverse collection of established datasets encompassing multiple sound domains, such as environmental audio, speech, music and many others (WU et al., 2023; MIAO et al., 2025).

The architecture of CLAP consists of two independent encoders: an audio encoder, based on HTSAT (Hierarchical Token-Semantic Audio Transformer) (CHEN et al., 2022) and PANN (Pretrained Audio Neural Networks) (KONG et al., 2020), and a text encoder, based on CLIP transformer (RADFORD et al., 2021), BERT (DEVLIN et al., 2019), and RoBERTa (LIU et al., 2019). These encoders extract high-level features from their respective inputs and project them into a shared latent space through a linear layer (RADFORD et al., 2021). The training objective is governed by a contrastive loss function that maximizes the cosine similarity between matched audio-text pairs and minimizes it between mismatched pairs within a batch.

This dual-encoder design enables the model to perform diverse tasks, including audio-text retrieval and zero-shot classification. CLAP can generalize to unseen sound categories by leveraging the semantic knowledge embedded in the text encoder, without requiring labels from a specific target domain. The robustness of the model relies on its extensive pre-training on a combination of large-scale datasets, most notably LAION-Audio-630K (WU et al., 2023) and AudioSet (GEMMEKE et al., 2017).

2.3.2. Text query creation

The formulation of textual queries can influence the performance of multimodal language-audio models, making prompt engineering crucial. Variations in query design can affect similarity scores and retrieval performance (MIAO et al., 2025). In this context, this work defined two experimental scenarios (A and B), each based on a different approach to query construction (structured vs descriptive prompts), to evaluate their impact on model performance. This design allows isolating the effect of linguistic abstraction on cross-modal retrieval.

2.3.2.1. Experimental Scenario A

In Experimental Scenario A, textual queries were constructed from acoustic features and metadata compiled in a spreadsheet for each acoustic recording. This metadata included the common name of the species, the sound type, and duration of the signals. The metadata were exported as a collection of dictionaries and stored in a structured *.json* file, which the retrieval pipeline subsequently parsed. During processing, the metadata file was read, and for each valid entry, a standardized English-language textual description was automatically generated within the code according to a fixed syntactic template:

“*The [common name] sound is a [sound type] type, with a duration of [duration] seconds.*”

This formulation was designed to provide the least amount of semantically informative content possible about the biological source and general acoustic characteristics of the signal, while avoiding overly descriptive or subjective language.

2.3.2.2. Experimental Scenario B

In Experimental scenario B, textual queries were created through auditory and visual inspection (spectrogram-based) of each audio recording. The aim was to incorporate perceptual descriptors commonly found in general acoustic and musical contexts, such as attributes related to timbre, texture and rhythm structure into the annotated sound type associated with each recording.

This approach was adopted to better align the textual queries with the training domain and the linguistic architecture of CLAP. Although CLAP was not specifically trained on bioacoustic datasets, it was exposed to diverse audio–text domains, including ambient sound effects, music and human speech. Thus, this descriptive language was used to represent fish signals in a form more compatible with those learned by the model.

Therefore, we transformed acoustic features into a standardized set of perceptual descriptors in order to ensure consistency and reproducibility. Each query was designed to capture at least one attribute from the following categories:

1. Temporal pattern: describes the rhythmic arrangement and duration of acoustic events (e.g., *“short pulses”*, *“rhythmic bursts”*, *“irregular intervals”*, or *“sustained signals”*).
2. Timbre and texture: maps the sound quality to common environmental or musical terms compatible with the training set of CLAP (e.g., *“sharp pops”*, *“dry knocks”*, *“metallic clicks”*, or *“raspy sounds”*).
3. Spectral dynamics and pitch: captures the behavior of frequency over time (e.g., *“low-frequency thumps”*, *“sliding pitch”*, or *“tonal boatwhistle-like sounds”*).

All queries were created manually and included as an additional column in the original metadata spreadsheet. Due to their descriptive and heterogeneous nature, the queries did not follow a fixed syntactic template. The retrieval pipeline read these entries directly and associated each query with its corresponding species and audio file identifier. Apart from the query construction procedure, all subsequent steps of embedding computation and similarity-based retrieval were identical to those in Experimental Scenario A.

2.3.3. Model implementation

We used the CLAP model to perform zero-shot classification on the constructed dataset, leveraging its shared latent space to assess the similarity between audio samples and textual descriptions. This process occurred in three main stages: *(1)* the extraction of high-dimensional embeddings for both modalities; *(2)* the computation of cosine similarity to retrieve acoustically compatible files; *(3)*

the evaluation of classification performance. The following subsections detail these stages.

2.3.3.1. Audio and text embeddings extraction

The CLAP model was initialized using the pre-trained weights from the LAION-Audio-630K checkpoint, which provides a robust latent space derived from over 600,000 audio-text pairs. The model was employed off-the-shelf (i.e., without fine tuning in our dataset). The architecture was configured without the fusion mechanism, using the HTSAT as the audio encoder and RoBERTa as the text encoder. This dual-encoder approach was maintained for the subsequent zero-shot classification tasks.

Text embeddings were generated in batch mode using the text encoder of the CLAP model. Given a set of *N* textual descriptions, the model produced fixed-length embedding vectors of dimension *D*, resulting in a text embedding matrix of size N_xD. All text embeddings were subsequently normalized, enabling the use of cosine similarity as a direct measure of semantic proximity. This normalization ensures that the inner product of vectors corresponds to cosine similarity by constraining the similarity values to the standard range of the cosine function.

Audio embeddings were extracted from the collection of recordings using the CLAP audio encoder and computed in batches. Each audio recording was mapped to the same *D* dimensional latent space as the textual descriptions, yielding an audio embedding matrix of size M_xD, where *M* denotes the number of audio files. As with the textual embeddings, all audio embeddings were normalized to ensure compatibility in cross-modal similarity computation.

2.3.3.2. Cosine similarity search

The similarity between textual descriptions and audio recordings was computed by constructing a similarity matrix based on the dot product between L2-normalized embeddings, which is equivalent to cosine similarity. Formally, the similarity matrix S was defined as:

$$S = E_{\text{text}} \cdot E_{\text{audio}}^T, \tag{1}$$

where $E_{\text{text}} \in \mathbb{R}^{N \times D}$ denotes the matrix of text embeddings and $E_{\text{audio}} \in \mathbb{R}^{M \times D}$ denotes the matrix of audio embeddings, both projected into a shared D-dimensional latent space. T represents the transpose operation required to compute the inner product between embeddings. Each element S_{ij} corresponds to the cosine similarity between the i-th textual query and the j-th audio recording. Following common practices in multimodal retrieval, similarity is computed as the dot product between normalized embeddings. While some works include an additional temperature scaling term (ELIZALDE et al., 2024), this factor is not required for ranking-based retrieval at inference time.

For each textual query, the Top-*K* most similar audio recordings were retrieved by ranking the similarity scores, with *K=5*. This procedure constitutes an N-to-M retrieval scenario, in which multiple textual queries are simultaneously compared against the entire set of available audio recordings.

The retrieval results were organized into a structured spreadsheet in which, for each textual query, the Top-*K* retrieved audio candidates, their similarity scores, and their respective retrieval ranks were recorded. Additionally, an automatic correspondence check was performed by comparing the retrieved audio filenames with the original audio associated with each textual description, allowing the identification of correct audio–text matches within the top-ranked candidates.

2.3.4. Performance evaluation

To address the specificities of fish sound retrieval as a downstream task for CLAP and to account for the complexities of evaluation across multiple taxonomic levels, we proposed a specialized evaluation framework. This approach characterizes model behavior in scenarios where zero-shot recognition capabilities are demonstrated. Since CLAP was not pretrained on bioacoustic data, focusing on successful retrievals helps reduce the impact of non-informative comparisons. Thus, three metrics were used for each taxonomic level (family, genus, and species), based on a total of N = 474 queries:

(A) Hit Rate@5

This is a binary success metric that determines if the target category is among the Top-*5* candidates suggested by the model. A "hit" is recorded when at least one audio file belonging to the target taxonomic group or sonotype appears in the ranking. It is defined as follows:

$$HR@5 = \frac{1}{N}\sum_{i=1}^{N} hit_i$$

where hit_i is 1 if a correct match exists in the Top-*5* results for query *i*, and 0 otherwise.

(B) Average Retrieval

We introduced the Average Retrieval metric to complement the Hit Rate@5. Average Retrieval quantifies whether the model retrieves multiple representatives of the same taxonomic category in the results. This is an adapted metric that accounts for all correct candidates identified in the top five, providing a broader measure of the consistency of the model in identifying relevant acoustic patterns across the dataset:

$$AR = \frac{\sum_{i-1}^{N} C_i}{N}$$

where C_i represents the total number of correct candidates identified within the Top-*5* results for query *i*.

(C) Mean Rank

To evaluate the positional precision of the model, the Mean Rank was calculated exclusively for successful queries (*S*). This metric reflects the average position (from 1 to 5) where a correct match was found within the Top-*5*. This metric was adapted from Mean Reciprocal Rank (MRR) to facilitate interpretation, as it provides directly the average position in the rank. A lower Mean Rank indicates higher model confidence in ranking the relevant audio closer to the first position:

$$MR = \frac{1}{S}\sum_{j=1}^{S} R_j$$

where R_j is the rank of the first correct candidate identified for a successful query j.

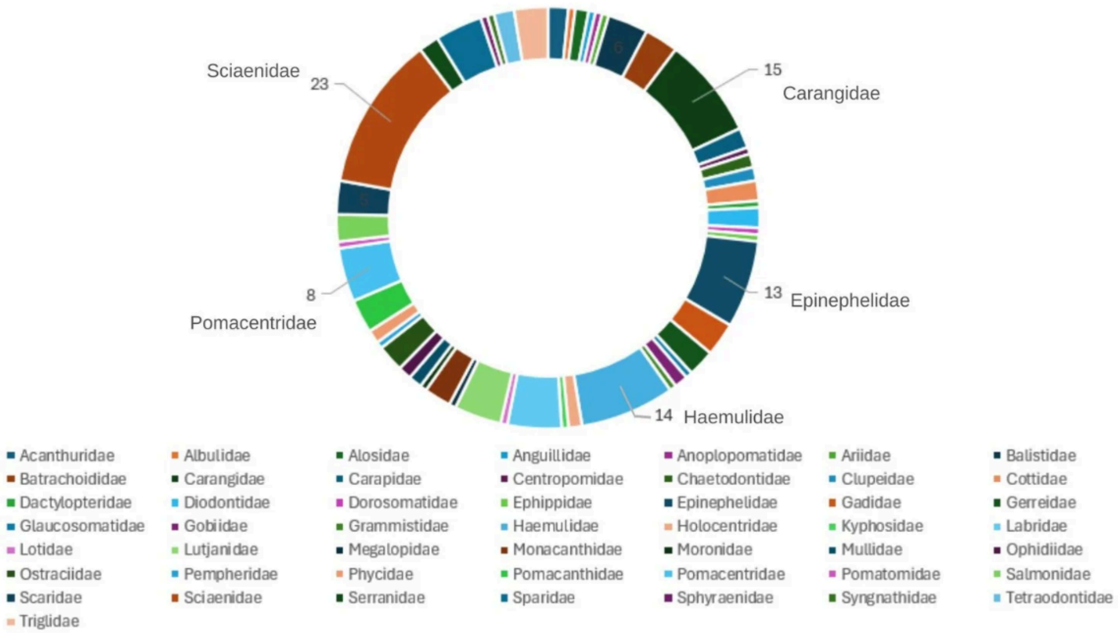
It should be noted that the sonotype level was assessed exclusively through Hit Rate@5 and Mean Rank. Since sonotype is not a taxonomic level and each one typically represents the most granular level in the dataset, the Average Retrieval metric becomes redundant in this context, as multiple correct candidates for a single sonotype are generally not expected within the Top-*5* results.

3. RESULTS

3.1. Exploratory data analysis

The resulting dataset comprises 474 audio samples corresponding to the sounds produced by 194 species, which are distributed across 121 genera and 50 families. The Sciaenidae was the most represented family in the dataset with 23 species. The Carangidae, Haemulidae, and Epinephelidae families followed with 15, 14, and 13 species, respectively. Pomacentridae and Labridae were each represented by eight species in the dataset (Figure 2). At the species level, the number of sonotypes per species ranged from one to thirteen (see Supplementary Material Table S1 for a summary of the dataset composition).

FIGURA 2 - Distribution of fish species by family in the dataset. The dataset includes species from 50 families, with an uneven taxonomic representation.



Among the sound libraries used to compile the dataset, the GSO Fish Sounds Collection (FISH; MOWBRAY, 1970) contributed the largest number of recordings (183), followed by FishBase (FROESE; PAULY, 2024) with 140 recordings (Table 1).

TABELA 1 - Distribution of sound files by sound libraries

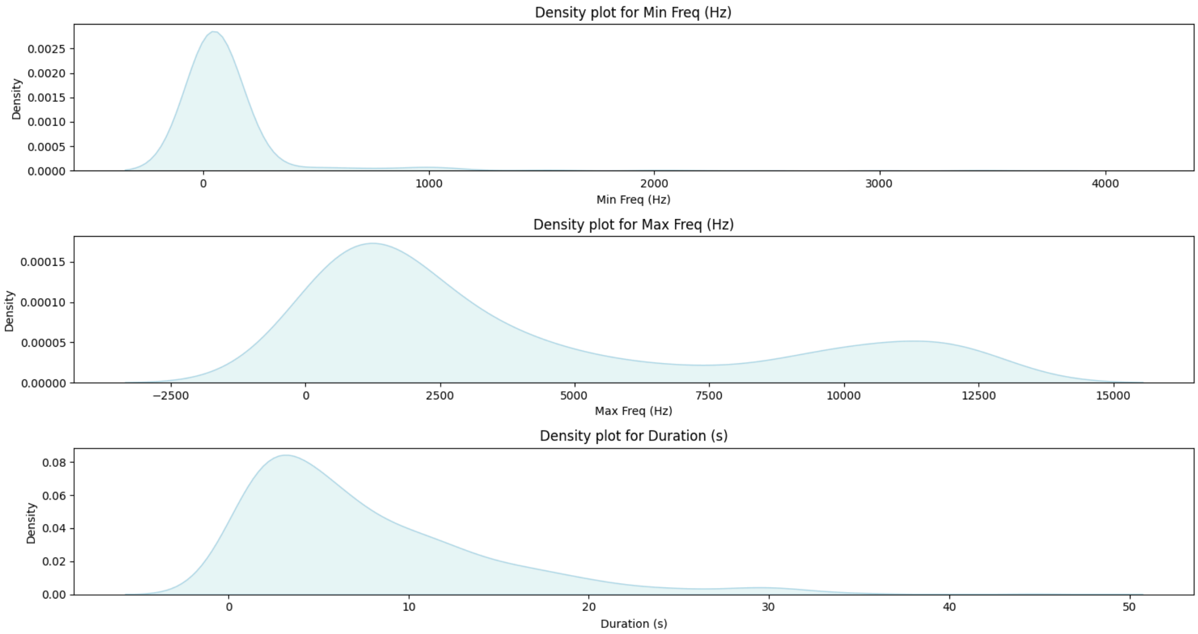
Library	Number of sound files
GSO Fish Sounds Collection	183
Fish Base	140
Fish Sounds	75
DOSITS	52
IEAPM	24

In terms of acoustic features, the majority of the sonotypes were concentrated in the low-frequency range (< 1 kHz). However, some species exhibited higher-frequency emissions, with maximum frequencies reaching 12 kHz.

The minimum frequency (Hz) and duration (s) are both skewed to the left, indicating that these two variables have discrepant values, with concentrations at lower frequencies (below 1 kHz) and shorter vocalization durations (less than 20 s). In contrast, the maximum frequency showed a pronounced concentration between

20 Hz and 2.5 kHz, as well as a secondary concentration in the 10-12 kHz range. Figure 3 shows the distributions of minimum and maximum frequencies, as well as sonotype durations, across the dataset.

FIGURA 3 - Kernel Density Estimation (KDE) plots showing the distributions of minimum frequency (Hz), maximum frequency (Hz), and duration for the sonotypes comprising the generated dataset.



3.2. Zero-shot classification of fish sounds

We evaluated the zero-shot retrieval performance of the CLAP model at different taxonomic resolutions and at the sonotype level. Two experimental setups were tested, differing in the formulation of text queries: Experimental Scenario A used queries derived from objective acoustic features, while Experimental Scenario B used queries incorporating perceptual descriptors. We assessed the performance of the model using cosine similarity and three metrics: hit rate@5, average retrieval, and mean rank. The results for these metrics are summarized in Tables 2 and 3.

Audio–Text Similarity Across Taxonomic Levels

We observed that Experimental Scenario B had higher similarity values across all taxonomic levels than Experimental Scenario A (Table 2). The highest average similarity was for the species level in Experimental Scenario B (0.525), followed by the genus level and sonotypes (0.522 and 0.519, respectively). Notably, the maximum similarity in Experimental Scenario B remained relatively higher across taxonomic levels (ranging from 0.636 to 0.690), but the minimum similarity values were substantially higher than in Experimental Scenario A for most taxonomic levels.

TABELA 2 - Cosine similarity values (minimum, maximum and average) obtained from fish sound classification using the CLAP model across two experimental scenarios (A and B) for the sonotype and taxonomic levels.

Cosine similarity								
Level	**Sonotype**		**Species**		**Genus**		**Family**	
Experimental scenario	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Minimum	0.298	0.432	0.267	0.364	0.267	0.330	0.267	0.285
Maximum	0.636	0.690	0.636	0.686	0.637	0.686	0.650	0.686
Average	0.444	0.519	0.400	0.525	0.441	0.522	0.497	0.514

Effectiveness and Consistency of Taxonomic Retrieval

The retrieval effectiveness followed a clear taxonomic trend, with hit rate@5 and average retrieval increasing at broader taxonomic levels. In general, Experimental Scenario B consistently outperformed Experimental Scenario A, with a localized exception for average retrieval at the genus level. The model demonstrated its best performance at the family level, reaching a hit rate of 44.00 and an average retrieval of 6.06 in Experimental Scenario B. In contrast, the model presented the lowest hit rate @ 5 values for sonotypes (2.53 and 4.22 for experimental scenarios A and B, respectively). At the species level, the hit rate increased from 7.73 in Experimental Scenario A to 11.34 in Experimental Scenario B (Table 3).

Ranking Precision of Correct Candidates

Regarding positional precision, the mean rank for successful hits remained remarkably consistent across all taxonomic levels and experimental scenarios, ranging from 2.50 to 3.10. This indicates that, when a hit occurs, the position among the Top-5 candidates tends to be between 2 and 3. Experimental Scenario B presented the best mean rank results for taxonomic levels, while Experimental

Scenario A showed the best result for the sonotype level. Table 3 shows the results for the retrieval performance across all taxonomic resolutions and sonotypes.

TABELA 3 - Performance metrics of the classification model evaluated at the sonotype level and at taxonomic levels (species, genus, and family) for fish sounds in Experimental Scenario A and Experimental Scenario B. The metrics include hit rate, average retrieval, and mean ranking.

Performance metrics								
Level	**Sonotype**		**Species**		**Genus**		**Family**	
Experimental Scenario	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Hit rate@5	2.53	4.22	7.73	11.34	15.70	19.10	36.00	44.00
Average retrieval	*	*	0.30	0.36	0.83	0.73	5.38	6.06
Mean ranking	2.50	2.95	2.90	2.79	3.08	2.79	3.10	2.93

* Average retrieval was not calculated for sonotype, as it is not a taxonomic level.

Taxonomic Distribution and Structure of Retrieval Results

The absolute number of correct identifications followed a consistent taxonomic trend. While identifications at the first level were limited to 12 and 20 sonotypes for experimental scenarios A and B, respectively, these counts increased substantially at higher ranks. Since sonotypes represent the primary acoustic unit rather than a formal taxonomic rank, the number of successful identifications at this level is reported solely as the count of correct candidates to avoid redundancy with taxonomic metrics.

Experimental Scenario B showed a consistent increase in the number of successfully identified taxa compared to Experimental Scenario A across all categories. Specifically, the number of correctly identified taxa increased from 15 species and 19 genera in Experimental Scenario A to 22 species and 23 genera in Experimental Scenario B. The highest absolute frequency of successful identifications occurred at the family level. where the number of correct candidates identified in Experimental Scenario B reached 303, compared to 269 in Experimental Scenario A. In Experimental Scenario B, the model correctly identified 22 out of 50 target families. Table 4 shows the total frequency of correct candidates and taxa identification at the sonotype, species, genus and family levels.

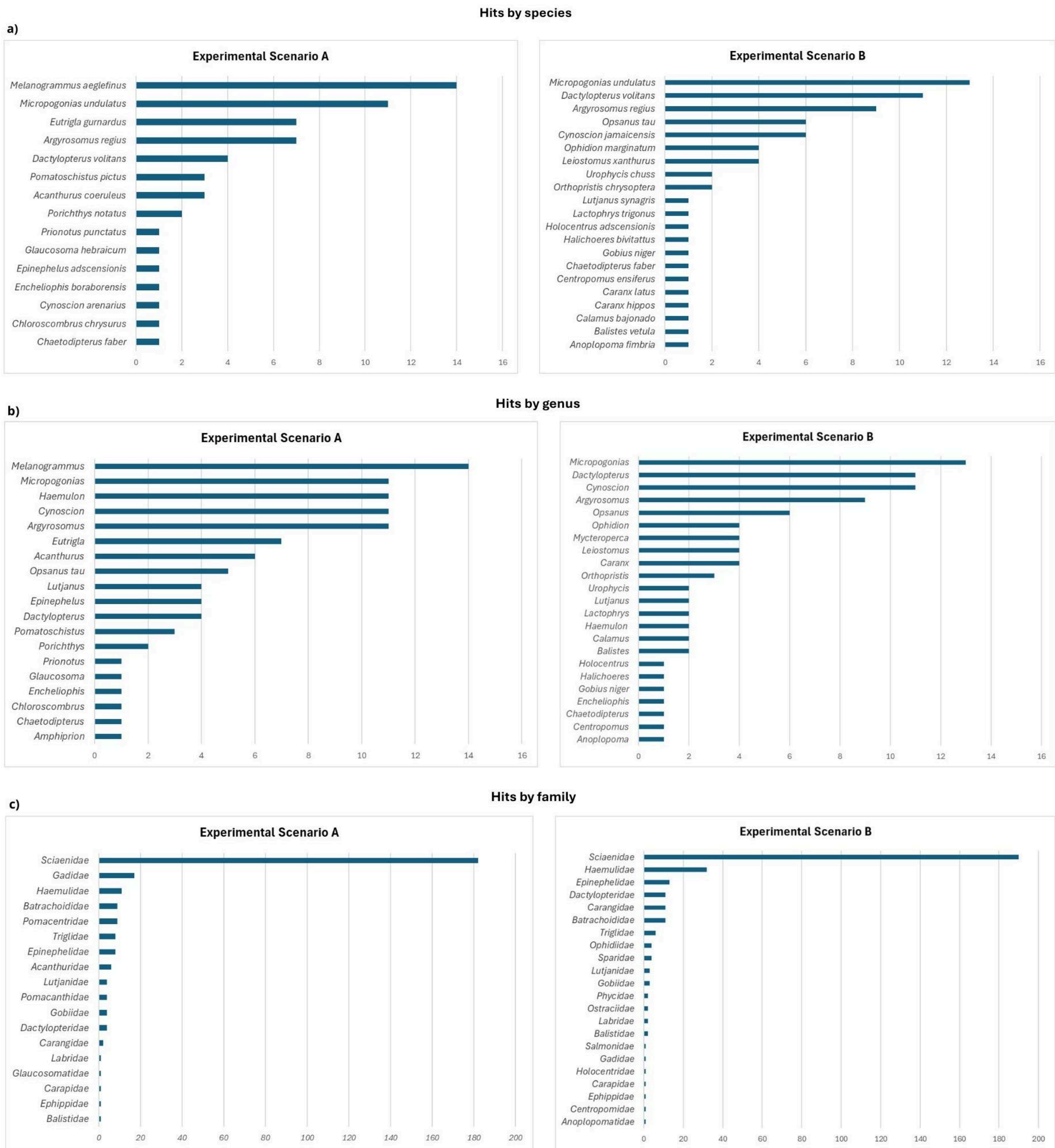
TABELA 4. Comparison of absolute retrieval counts between experimental scenarios A and B.

Absolute Retrieval								
Level	**Sonotype**		**Species**		**Genus**		**Family**	
Experimental Scenario	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Total	474		194		121		50	
Correct candidates	12	20	58	69	100	88	269	303
Correct taxa	-	-	15	22	19	23	18	22

Beyond the overall performance metrics, a visual analysis of retrieval frequency (Figure 4) reveals which specific taxa drove these results. The distribution of hits across different taxonomic resolutions shows that retrieval success was concentrated among certain biological groups. At the species level, *Melanogrammus aeglefinus* (Gadidae) and *Micropogonias undulatus* (Sciaenidae) presented the highest number of hits in Experimental Scenario A. However, Experimental Scenario B significantly increased the retrieval frequency of sciaenids, such as *Micropogonias undulatus*, *Argyrosomus regius, Cynoscion jamaicensis* and *Leiostomus xanthurus*. This trend is also evident at the genus level, where *Micropogonias* and *Cynoscion* emerged as the dominant groups in the second experimental scenario.

The highest concentration of hits occurred at the family level. In both experimental scenarios, Sciaenidae consistently dominated the retrieval results, with a higher absolute number of identifications in Experimental Scenario B. This family was the main contributor, accounting for 182 out of 269 total family-level identifications in Experimental Scenario A and 190 out of 303 in Experimental Scenario B. Consequently, although the absolute number of Sciaenidae hits increased in Experimental Scenario B, its relative contribution decreased slightly from 68% to 63%. Other families, such as the Batrachoididae, Haemulidae, and Epinephelidae, also exhibited significant retrieval counts in both experimental scenarios. Notably, the latter two families demonstrated substantial increases in Experimental Scenario B. These results are summarized in Figure 4, which shows the absolute number of correct candidate retrievals across species, genus, and family levels for Experimental scenarios A and B.

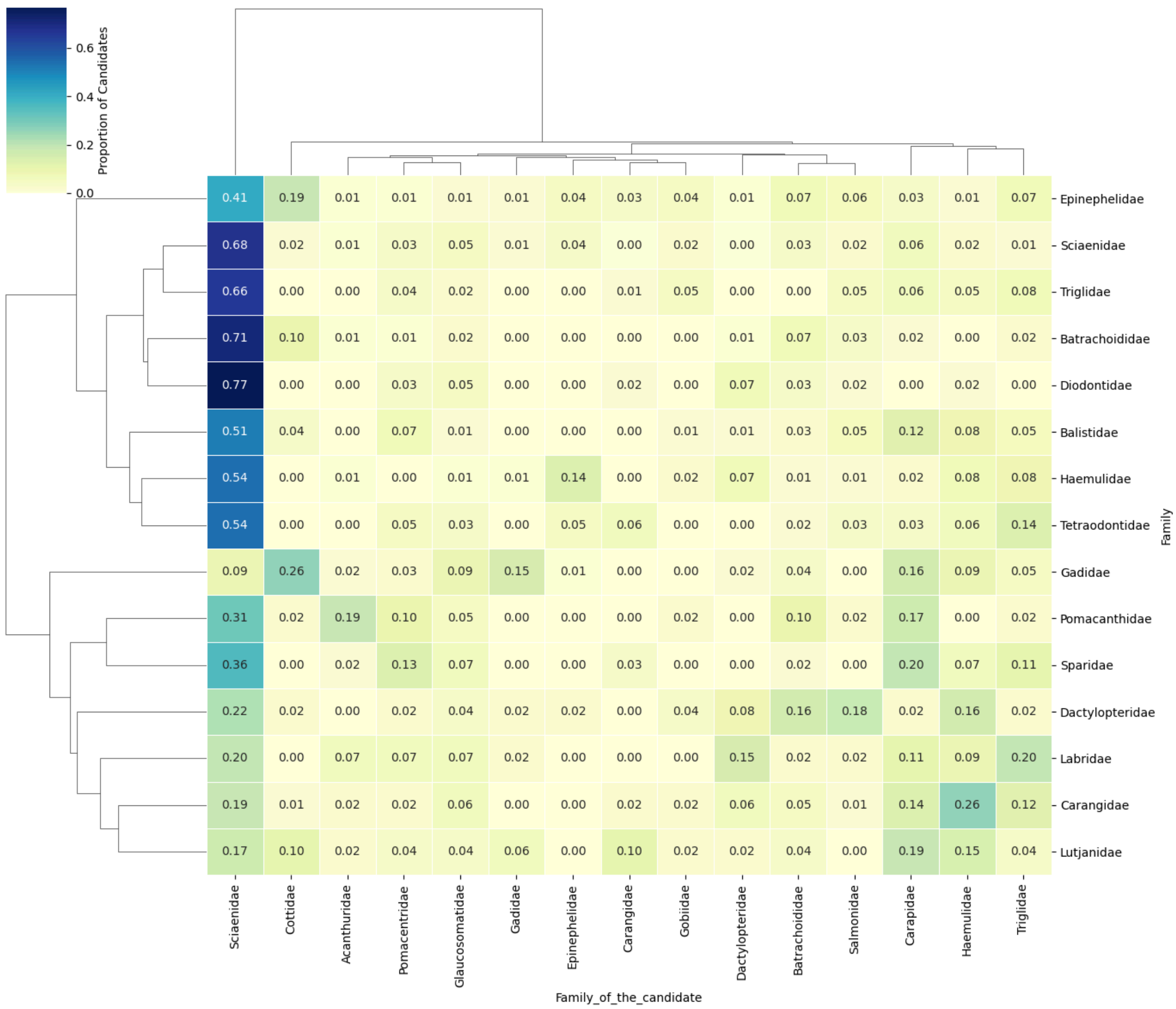
FIGURA 4 - Absolute number of correct candidate retrievals for Experimental scenarios A and B across taxonomic levels. The charts highlight the taxa with the highest retrieval frequency. Horizontal bar charts illustrate the total number of correct candidate identifications for (a) species, (b) genera, and (c) families.



While the analysis of absolute hit counts provides an overview of retrieval performance across taxonomic levels, it does not capture how retrieval errors and similarities are distributed among families. During an initial inspection of the retrieval results, we observed that, in most cases, the candidate families were not randomly distributed. Instead, specific families repeatedly emerged as candidates. To investigate these non-random associations and the nature of taxonomic misidentifications at the family level, we examined the distribution of candidate families retrieved for each original family using normalized heatmaps.

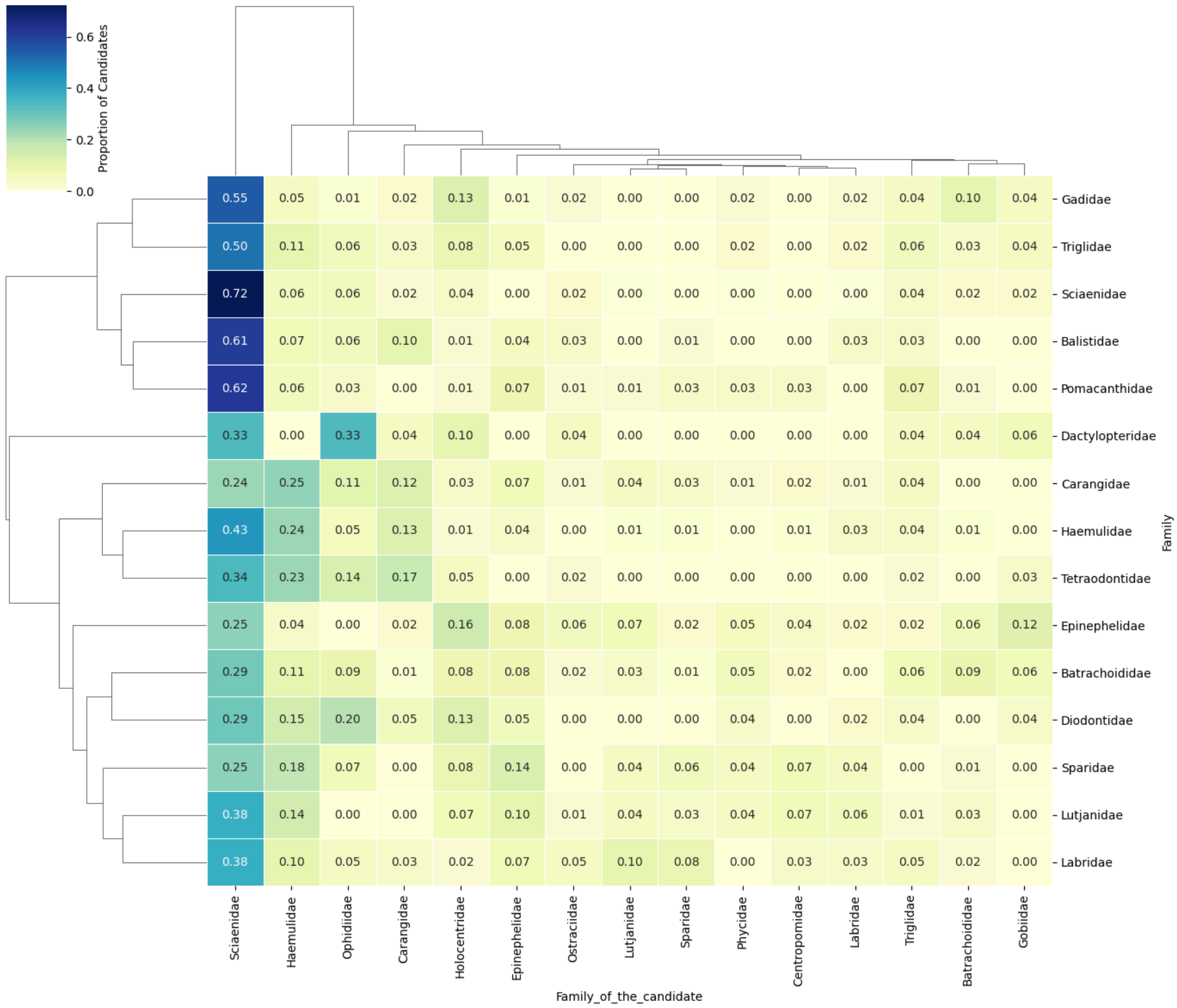
Figures 5 and 6 summarize the family level retrieval patterns obtained in experimental scenarios A and B, respectively. To enhance interpretability and reduce sparsity effects associated with rare families, the analysis was restricted to the 15 most frequent original families and the 15 most frequent candidate families. The resulting heatmaps depict row-normalized proportions of candidate families retrieved for each original family.

FIGURA 5 - Hierarchically clustered heatmap of family-level retrieval associations for Experimental Scenario A.



In Experimental Scenario A (Figure 5), the matrix shows a concentration of retrievals involving Sciaenidae, which exhibited a high intra-family association (0.68) and also emerged as a frequent candidate for families such as Diodontidae (0.77) and Batrachoididae (0.71). However, additional structured associations are also observed among other families. For example, Batrachoididae, Haemulidae, and Epinephelidae exhibit relationships and form groups in the hierarchical clustering, indicating non-random inter-family retrieval patterns beyond a single dominant family.

FIGURA 6 - Hierarchically clustered heatmap of family-level retrieval associations for Experimental Scenario B.



In contrast, the heatmap for Experimental Scenario B (Figure 6) exhibits a more distributed pattern of higher values along the matrix. While Sciaenidae remains prominent, increased proportions are observed for other families, including Dactylopteridae (0.33). The hierarchical clustering further reflects these changes, revealing more compact and distinct family groupings. Notably, Haemulidae, Carangidae, and Tetraodontidae form clusters in Experimental Scenario B, a structure that is less apparent in Experimental Scenario A.

4. DISCUSSION AND CONCLUSIONS

The application of audio–text embeddings to bioacoustic classification remains a relatively unexplored research direction, particularly in domains characterized by high taxonomic diversity and acoustic variability, such as fish sounds. A previous study has demonstrated the potential of multimodal models, including CLAP, for recognizing broad animal sound categories (e.g., frogs, birds, whales, and meerkats) (MIAO et al., 2025). However, fine-grained taxonomic recognition has been identified as a persistent challenge for these architectures. In this context, the results of the present study indicate that CLAP and derived/similar audio–text embedding models can retrieve biologically meaningful acoustic patterns at multiple taxonomic resolutions, including family, genus, and species levels. Notably, isolated correct retrievals at the sonotype level further suggest that the model is capable of capturing highly specific acoustic signatures, even in the absence of domain-specific training and under substantial inter- and intra-species variability. These findings position multimodal audio–text embeddings as a promising complementary approach for fish bioacoustics, particularly in data-scarce scenarios.

Similarity scores (Table 2) and retrieval performance metrics (hit rate@5, average retrieval, and mean rank) (Table 3) were consistently higher for descriptive queries (Experimental Scenario B) than for structured queries (Experimental Scenario A). This improvement can be primarily attributed to a richer semantic alignment between the textual queries and the representations learned by CLAP. While autoencoder-based approaches can enhance the relevance of vocalization representation (BEST et al., 2023), zero-shot audio–text models rely on learning the relationships between the acoustic semantics and language semantics. This enables flexibility and generalization (ELIZALDE et al., 2022). In addition, variations in text prompts can significantly affect the model prediction performance (MIAO et al., 2025). While the queries in Experimental Scenario A were limited to minimal metadata, such as species common name, sound type, and duration, this formulation provided limited discriminative acoustic information, resulting in text embeddings with constrained capacity. In contrast, Experimental Scenario B incorporated perceptual descriptors related to temporal patterns, timbre, texture, and spectral dynamics. These descriptors are more closely associated with the audio–text correspondences learned during the model pretraining on diverse sound domains.

The results showed a clear trend of improved performance at broader taxonomic levels (from sonotype to family). This progression is observed in the similarity (Table 2), the performance metrics (Table 3) and the absolute values of retrieval (Table 4). This progression is expected, as broader taxonomic levels, such as family, aggregate multiple lower-level units, thereby increasing the probability of hits within the Top-*5* candidates. However, many hits occurred exclusively at the family level. This suggests that the model captures acoustic patterns that are shared by related taxa more effectively than the subtle distinctions for the species-specific sounds. These shared acoustic patterns are possibly more distinct in the embedding space than the fine-grained variations between individual species or sonotypes.

In addition to the taxonomic hierarchy, retrieval patterns were shaped by query specifics and sample distribution within the dataset. A localized exception was observed at the genus level, with a higher average retrieval in Experimental Scenario A, which can be explained by changes in the distribution of hits within the Top-*5* candidates. This suggests that while the queries used in Experimental Scenario B enhanced the ability of the model to correctly identify a higher number of genera, the queries used in Experimental Scenario A tended to group more candidates of the same genus within the Top-*5*. An example of this behavior is the genus *Melanogrammus*, which is represented by only one species in the dataset (*Melanogrammus aeglefinus*), presented 14 hits in Experimental Scenario A. This can be attributed to the large number of samples for this species (12 sonotypes) and the lack of specificity of the queries in Experimental Scenario A. Another observation regarding the performance metrics was the inverse relationship between the correct identifications and the mean rank. As the model correctly identified more candidates within a specific group, the mean rank decreased. This can be explained by the fact that when many hits occurred for certain families, the entire Top-*5* rank became populated with a higher number of candidates. In other words, hits occurred for all five candidates in most cases, especially at the family level.

It is important to note that the performance metrics observed must be interpreted considering the CLAP training domain and the nature of the acoustic data considered in this study. The dataset is composed of fish sounds, which are unfamiliar to the original training domain of the model. Given that the CLAP model was not trained in the bioacoustic domain (ROBINSON et al., 2024; MIAO et al., 2025), the observed performance may reflect the domain mismatch between the training data and fish signals. The EDA performed in this study revealed that the dataset is acoustically and taxonomically heterogeneous (Figure 2 and Supplementary Material Table S1). It comprises a wide range of families, genera, and species, as well as substantial variability in the frequency content and duration of the sonotypes (Figure 3). The predominance of low-frequency sounds, coupled with uneven representation across taxonomic levels, poses additional challenges for similarity-based retrieval.

In addition to the performance metrics, the distribution of retrieval patterns provides a qualitative view of the observed rank variations. The heatmaps in Figures 5 and 6 suggest that groups with similar acoustic characteristics converge within the embedding space, suggesting the presence of structured patterns of association between families. This behavior indicates that the model generalizes over shared acoustic structures, reflecting a learned semantic acoustic space. This pattern is particularly clear in the results for the Sciaenidae family.

At the family level, Sciaenidae emerged as the dominant group in the retrieval results in both experimental scenarios. In Experimental Scenario A, this family accounted for 182 hits, corresponding to 68% of the total hits and, in Experiment B, it reached 190 hits, corresponding to 63% of the total hits. Although the absolute number of hits for Sciaenidae has increased in Experimental Scenario B, this increase occurred alongside a broader rise in total identifications at the family level, indicating a diversification of retrieved families rather than a reduction in Sciaenidae detections. This group was indeed the most represented in the dataset, with 23 species. However, the success of the retrieval cannot be attributed solely to this fact. This family was characterized by being the only one with at least one correct candidate in the Top-*5* for all the 23 species except one (*Menticirrhus littoralis* and *Larimus breviceps* in experimental scenarios A and B, respectively). In addition, this family had hits at the sonotype, genus, and species levels, often occupying all five positions in the ranking.

This level of consistency was not observed for the other families in the dataset, including the most represented families: Carangidae, Haemulidae, Epinephelidae, and Pomacentridae. For example, Carangidae was the second family with the highest number of species in the dataset (15 species). However, this group had only two hits in Experimental Scenario A and 11 in Experimental Scenario B, representing 0.7% and 3.6% of the correct identifications, respectively.

The discrepancy in the number of hits for Sciaenidae compared to other well-represented families suggests that the model captured acoustic patterns of Sciaenidae. Commonly referred to as "croakers" or "drums" due to their characteristic sounds, these species are among the most studied regarding bioacoustics (TELLECHEA et al., 2025). These sounds include low-frequency drums, grunts, thumps and knocks (RAMCHARITAR, 2006; PARMENTIER, 2017; VIEIRA et al., 2019) produced by the contraction of sonic muscles attached to the swim bladder (BORIE et al., 2014; FINE, 2023). This indicates a more effective alignment between the latent features of Sciaenidae sounds and their respective textual descriptions in the embedding space.

One hypothesis is that this alignment is due to the characteristics of CLAP pretraining datasets (GEMMEKE et al., 2017; WU et al., 2023), which include music. Specifically, the exposure of the model to rhythmic and periodic structures in music may have optimized its ability to recognize the temporal periodicity of Sciaenid drumming patterns as percussive events. Despite the large amount of sound examples of Carangidae in the dataset, their specific acoustic patterns (e.g. higher frequency bursts) (FISH; MOWBRAY, 1970; CHANDA et al., 2020), which are less temporally regular and may lack the temporal distinctiveness that the model is optimized to recognize. Besides, the HTSAT encoder uses blocks of swin transformer (LIU et al., 2021; WU, 2023), which allows CLAP to process spectrograms in a more structured manner, leveraging the hierarchical tokenization and self-attention mechanisms (RENTOULA, 2025). Unlike traditional methods that focus on isolated acoustic patterns, this approach can extract a global feature representation, which captures rhythmic patterns and spectral texture, which are characteristic of specific taxa.

Therefore, the comparatively superior performance of CLAP for Sciaenidae may be partly explained by the distinctive and well-structured acoustic properties of their calls. Their vocalizations present common characteristics, such as the rapid decay in the amplitude envelope of pulses, low intraspecific variability in the dominant frequency and silent periods between pulses (PARMENTIER et al., 2018). These characteristics also allowed CLAP to differentiate between Sciaenidae and other families that produce similar sound types as grunts, knocks, boops and thumps, such as Haemulidae (MILLOT et al., 2021), Gadidae (HAWKINS; AMORIM, 2000), and Batrachoididae (STAATERMAN et al., 2018).

Previous studies used ML and DL to detect and classify sounds of Sciaenidae species using different techniques and feature sets (RUIZ-BLAIS et al., 2012; HARAKAWA et al., 2018; VIEIRA et al., 2019; MONCZACK et al., 2019). While these studies relied on predefined acoustic features and species-specific modeling, the present work advances this line of research by incorporating natural language–based textual descriptions jointly with audio embeddings. This approach enables the recognition and retrieval of sciaenid sounds through semantically meaningful descriptions, reducing dependence on narrowly defined acoustic features and offering a more flexible and extensible paradigm for bioacoustic analysis, particularly in contexts where annotated data are scarce or heterogeneous.

Sciaenids play key roles in structuring the dynamics of coastal food webs and are strongly associated with nearshore habitats that function as nursery, feeding and spawning grounds, making them a critical group for ecosystem-level monitoring (CHAO et al., 2015). Additionally, the family includes several commercially valuable species that sustain artisanal and industrial fisheries (CHAO et al., 2015) and are subject to intense fish exploitation pressure, often combined with habitat degradation (HAIMOVICI et al., 2016; WHITFIELD; GRIFFITHS, 2025). Many sciaenids are known for producing species- and context-specific sounds, especially during reproductive contexts, which make acoustic signals a direct proxy for biologically meaningful processes (RAMCHARITAR et al., 2006). In this context, the model's ability to capture consistent acoustic patterns within this family, even without domain-specific bioacoustic training, highlights the potential of multimodal and zero-shot approaches to support scalable and taxonomically informed monitoring tools. Such frameworks could contribute to tracking population dynamics, reproductive activity and habitat use of ecologically and economically important species, offering valuable support for conservation and fisheries management.

While these findings demonstrate the potential of multimodal embeddings, they also highlight the inherent challenges of applying them to the bioacoustics domain. The CLAP model was primarily pretrained on general audio datasets and lacks explicit exposure to the specificities of biological sounds. Consequently, the observed alignment likely relies on acoustic structures and descriptors more commonly found in speech and music. This reliance may constrain the model performance when applied to fish sounds, which are often low-frequency and highly diverse (LADICH, 2014). In addition, the dataset used in this study presents inherent complexity, including the diversity of species and sonotypes (194 and 474, respectively), as well as an imbalanced distribution of samples across taxa. These factors posed challenges to the proposed zero-shot framework, by limiting the capacity of the model to consistently retrieve fine-grained distinctions.

On the other hand, the model correctly identified 22 out of 50 target families and 20 sonotypes, which are the most granular level in this study. Although the use of embeddings for bioacoustics remains constrained without fine tuning (CHEN; YANG, 2025), the framework used in this study could reach the sonotype level, which is a challenging task for a nonpretrained model in such a specific domain as the fish sounds. In addition, the zero-shot classification demonstrated the capacity of the model to retrieve specific sonotypes containing a single sample in the dataset. This indicates the ability of the model to generalize even for rare data and suggests that the retrieval success is influenced more by the quality of the acoustic-semantic alignment than by the quantity of samples. These factors contextualize the proposed model as a preliminary step in transitioning traditional fish sound classification to a natural language-driven approach, as this field is characterized by the lack of labeled data for training classical AI frameworks (PARSONS et al., 2022).

To improve the effectiveness of multimodal models in the bioacoustics domain, future work should explore prompt engineering optimization by identifying textual descriptions that better represent the specificities of fish sounds, as the query modulation can impact the performance of the model. As a strategy to refine the retrieval outcomes, the development of algorithms to integrate automated search in global taxonomic databases, such as FishBase and WoRMS are encouraged. This could allow for using biological and geographical filters to exclude unlikely candidates within the Top-*5*. This integration could allow cross-validate acoustic embeddings with metadata, such as habitat and distribution, narrowing the space to plausible species and increasing the classification reliability.

5. REFERENCES

AUDACITY TEAM. *Audacity®: free audio editor and recorder* [software]. Versão 3.7.1. 2024.

BARROSO, V. R.; XAVIER, F. C.; FERREIRA, C. E. L. Applications of machine learning to identify and characterize the sounds produced by fish. *ICES Journal of Marine Science*, v. 80, p. 1854–1867, 2023.

BARROSO, V. R.; LESSA, A. A.; FERREIRA, C. E. L.; XAVIER, F. C. Optimal feature selection and model explanation for reef fish sound classification. *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 380, n. 1928, p. 20240055, 2025.

BEST, P.; PARIS, S.; GLOTIN, H.; MARXER, R. Deep audio embeddings for vocalisation clustering. *PLOS ONE*, v. 18, n. 7, p. e0283396, 2023.

BIANCO, M. J.; GERSTOFT, P.; TRAER, J.; OZANICH, E.; ROCH, M. A.; GANNOT, S.; DELEDALLE, C. A. Machine learning in acoustics: theory and applications. *The Journal of the Acoustical Society of America*, v. 146, p. 3590–3628, 2019.

BORIE, A.; MOK, H. K.; CHAO, N. L.; FINE, M. L. Spatiotemporal variability and sound characterization in Silver Croaker *Plagioscion squamosissimus* (Sciaenidae) in the Central Amazon. *PloS one*, v. 9, n. 8, p. e99326, 2014.

CHANDA, K.; CHAKRABORTY, B.; VARDHAN, Y. V.; GRACIAS, D.; MAHANTY, M. M.; LATHA, G. et al. Passive acoustic metrics to understand shallow water biodiversity off Malvan area in the west coast of India. *Indian J. Geo-Mar. Sci.*, v. 49, p. 527–536, 2020.

CHAO, N. L.; FRÉDOU, F. L.; HAIMOVICI, M.; PERES, M. B.; POLIDORO, B.; RASEIRA, M. et al. A popular and potentially sustainable fishery resource under pressure – extinction risk and conservation of Brazilian Sciaenidae (Teleostei: Perciformes). *Global Ecology and Conservation*, v. 4, p. 117–126, 2015.

CHEN, K.; DU, X.; ZHU, B.; MA, Z.; BERG-KIRKPATRICK, T.; DUBNOV, S. Hts-at: a hierarchical token-semantic audio transformer for sound classification and detection. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2022. *Proceedings…* Piscataway: IEEE, 2022. p. 646–650.

CHEN, C.; YANG, Z. No free lunch from audio pretraining in bioacoustics: a benchmark study of embeddings. *arXiv preprint*, 2025. Disponível em: https://arxiv.org/abs/2508.10230.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. Bert: pre-training of deep bidirectional transformers for language understanding. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2019. *Proceedings…* 2019. p. 4171–4186.

DISCOVERY OF SOUND IN THE SEA (DOSITS). *Audio gallery – fishes*. Disponível em: https://dosits.org/galleries/audio-gallery/fishes/.

ELIZALDE, B.; DESHMUKH, S.; AL ISMAIL, M.; WANG, H. Clap: learning audio concepts from natural language supervision. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2023. *Proceedings…* Piscataway: IEEE, 2023. p. 1–5.

ELIZALDE, B.; DESHMUKH, S.; WANG, H. Natural language supervision for general-purpose audio representations. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2024. *Proceedings…* Piscataway: IEEE, 2024. p. 336–340.

FINE, M. L. A personal account of work on anatomy, physiology, and behavior of fish sound production. *The Journal of the Acoustical Society of America*, v. 154, n. 2, p. 902–917, 2023.

FISH, M. P.; MOWBRAY, W. H. *Sounds of western North Atlantic fishes: a reference file of biological underwater sounds*. Baltimore: The John Hopkins Press, 1970.

FROESE, R.; PAULY, D. (ed.). *FishBase*. 2024. Disponível em: https://www.fishbase.org. Versão 10/2024. Acesso em: nov. 2024.

GANNON, D. P. Passive acoustic techniques in fisheries science: a review and prospectus. *Transactions of the American Fisheries Society*, v. 137, n. 2, p. 638–656, 2008.

GEMMEKE, J. F.; ELLIS, D. P.; FREEDMAN, D.; JANSEN, A.; LAWRENCE, W.; MOORE, R. C. et al. Audio set: an ontology and human-labeled dataset for audio events. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2017. *Proceedings…* Piscataway: IEEE, 2017. p. 776–780.

GOODWIN, M.; HALVORSEN, K. T.; JIAO, L.; KNAUSGÅRD, K. M.; MARTIN, A. H.; MOYANO, M.; OOMEN, R. A. et al. Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. *ICES Journal of Marine Science*, v. 79, p. 319–336, 2022.

HAIMOVICI, M.; CARDOSO, L. G.; UNPIERRE, R. G. Stocks and management units of *Micropogonias furnieri* (Desmarest, 1823) in southwestern Atlantic. *Latin American Journal of Aquatic Research*, v. 44, n. 5, p. 1080–1095, 2016.

HARAKAWA, R.; OGAWA, T.; HASEYAMA, M.; AKAMATSU, T. Automatic detection of fish sounds based on multi-stage classification including logistic regression via adaptive feature weighting. *The Journal of the Acoustical Society of America*, v. 144, n. 5, p. 2709–2718, 2018.

HAWKINS, A. D.; AMORIM, M. C. P. Spawning sounds of the male haddock, *Melanogrammus aeglefinus*. *Environmental Biology of Fishes*, v. 59, n. 1, p. 29–41, 2000.

HILDEBRAND, J. A. Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series*, v. 395, p. 5–20, 2009.

KASUMYAN, A. O. Acoustic signaling in fish. *Journal of Ichthyology*, v. 49, p. 963–1020, 2009.

KONG, Q.; CAO, Y.; IQBAL, T.; WANG, Y.; WANG, W.; PLUMBLEY, M. D. Panns: large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 28, p. 2880–2894, 2020.

LADICH, F. Fish bioacoustics. *Current opinion in neurobiology*, v. 28, p. 121–127, 2014.

LANCASTER, D.; MOUY, X.; HAGGARTY, D.; JUANES, F. Knock knock, who's there? Identifying wild species-specific fish sounds with passive acoustic localization and random forest models. *Journal of Fish Biology*, 2025.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D. et al. Roberta: a robustly optimized BERT pretraining approach. *arXiv preprint*, 2019. Disponível em: https://arxiv.org/abs/1907.11692.

LIU, Z.; LIN, Y.; CAO, Y. et al. Swin transformer: hierarchical vision transformer using shifted windows. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2021. *Proceedings…* 2021.

LOOBY, A.; VELA, S.; COX, K.; RIERA, A.; DAVIES, H. L.; SPRIEL, B.; MURCHY, K.; BRAVO, S.; ROUNTREE, R.; JUANES, F.; REYNOLDS, L. K.; MARTIN, C. W. *FishSounds*. Versão 3.1. 2024. Disponível em: http://www.fishsounds.net.

MIAO, Z.; ELIZALDE, B.; DESHMUKH, S.; KITZES, J.; WANG, H.; DODHIA, R.; FERRES, J. L. Multi-modal language models in bioacoustics with zero-shot transfer: a case study. *Scientific Reports*, v. 15, n. 1, p. 7242, 2025.

MILLOT, M.; BERTUCCI, F.; LECCHINI, D.; SMEETS, S.; RENÉ-TROUILLEFOU, M.; PARMENTIER, E. Characteristics of sound production and associated pharyngeal jaws in the tomtate grunt *Haemulon aurolineatum* (Cuvier, 1830) in Caribbean reefs. *Belgian Journal of Zoology*, v. 151, 2021.

MONCZAK, A.; JI, Y.; SOUEIDAN, J.; MONTIE, E. W. Automatic detection, classification, and quantification of sciaenid fish calls in an estuarine soundscape in the Southeast United States. *PloS one*, v. 14, n. 1, p. e0209914, 2019.

PARMENTIER, E.; DI IORIO, L.; PICCIULIN, M.; MALAVASI, S.; LAGARDÈRE, J. P.; BERTUCCI, F. Consistency of spatiotemporal sound features supports the use of passive acoustics for long-term monitoring. *Animal Conservation*, v. 21, n. 3, p. 211–220, 2018.

PARSONS, M. J.; LIN, T. H.; MOONEY, T. A.; ERBE, C.; JUANES, F.; LAMMERS, M.; LI, S. et al. Sounding the call for a global library of underwater biological sounds. *Frontiers in Ecology and Evolution*, v. 10, p. 39, 2022.

RADFORD, A.; KIM, J. W.; HALLACY, C.; RAMESH, A.; GOH, G.; AGARWAL, S. et al. Learning transferable visual models from natural language supervision. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 2021. *Proceedings…* PMLR, 2021. p. 8748–8763.

RAMCHARITAR, J.; GANNON, D. P.; POPPER, A. N. Bioacoustics of fishes of the family Sciaenidae (croakers and drums). *Transactions of the American Fisheries Society*, v. 135, n. 5, p. 1409–1431, 2006.

K. LISA YANG CENTER FOR CONSERVATION BIOACOUSTICS AT THE CORNELL LAB OF ORNITHOLOGY. *Raven Pro: interactive sound analysis software*. Versão 1.6 [software]. Ithaca: The Cornell Lab of Ornithology, 2023. Disponível em: https://www.ravensoundsoftware.com/.

RENTOULA, V. *Automatic music captioning*. 2025. Dissertação (Master's thesis) – University of Pireu, 2025.

RICE, A. N.; FARINA, S. C.; MAKOWSKI, A. J.; KAATZ, I. M.; LOBEL, P. S.; BEMIS, W. E.; BASS, A. H. Evolutionary patterns in sound production across fishes. *Ichthyology & Herpetology*, v. 110, n. 1, p. 1–12, 2022.

ROBINSON, D.; MIRON, M.; HAGIWARA, M.; WECK, B.; KEEN, S.; ALIZADEH, M. et al. Naturelm-audio: an audio-language foundation model for bioacoustics. *arXiv preprint*, 2024. Disponível em: https://arxiv.org/abs/2411.07186.

RUIZ-BLAIS, S.; RIVERA-CHAVARRIA, M. R.; CAMACHO, A. Autonomous detection of neotropical sciaenid fishes. In: PROCEEDINGS OF MEETINGS ON ACOUSTICS, 2012. Melville: Acoustical Society of America, 2012. v. 18, n. 1, p. 010001.

SANTOS, S.; DE FÁTIMA GOMES, M.; DOS SANTOS FERREIRA, A. R.; SAMPAIO, I.; SCHNEIDER, H. Molecular phylogeny of the western South Atlantic Sciaenidae based on mitochondrial and nuclear data. *Molecular phylogenetics and evolution*, v. 66, n. 1, p. 423–428, 2013.

STAATERMAN, E.; BRANDL, S. J.; HAUER, M.; CASEY, J. M.; GALLAGHER, A. J.; RICE, A. N. Individual voices in a cluttered soundscape: acoustic ecology of the Bocon toadfish, *Amphichthys cryptocentrus*. *Environmental Biology of Fishes*, v. 101, n. 6, p. 979–995, 2018.

TELLECHEA, J. S.; MOK, H. K.; FINE, M. L. Characterizing dual-pulse calls in five Sciaenid species. *Scientific Reports*, v. 15, n. 1, p. 42442, 2025.

VIEIRA, M.; PEREIRA, B. P.; POUSÃO-FERREIRA, P.; FONSECA, P. J.; AMORIM, M. C. P. Seasonal variation of captive meagre acoustic signalling: a manual and automatic recognition approach. *Fishes*, v. 4, n. 2, p. 28, 2019.

WHITFIELD, A. K.; GRIFFITHS, M. H. Biology, ecology and conservation of the IUCN endangered sciaenid *Argyrosomus japonicus* in global estuarine and marine waters. *Journal of Fish Biology*, v. 107, n. 5, p. 1481–1499, 2025.

WU, Y.; CHEN, K.; ZHANG, T.; HUI, Y.; BERG-KIRKPATRICK, T.; DUBNOV, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2023. *Proceedings…* Piscataway: IEEE, 2023. p. 1–5.

XAVIER, F. C. *Bioacoustic signature of rocky shores: characterization, modeling, and biotechnological applications*. 2021. Tese (Doutorado) – Universidade Federal Fluminense, Niterói, 2021.

ZHANG, C.; PENG, B.; SUN, X.; NIU, Q.; LIU, J.; CHEN, K. et al. From word vectors to multimodal embeddings: techniques, applications, and future directions for large language models. *arXiv preprint*, 2024. Disponível em: https://arxiv.org/abs/2411.05036.

5. CONSIDERAÇÕES FINAIS

A presente tese investigou a evolução e a aplicação de técnicas de Inteligência Artificial (IA) no campo da bioacústica de peixes marinhos, abordando desde a fundamentação teórica até o desenvolvimento de métodos de fronteira para a classificação automática desses sinais acústicos. Através de uma abordagem plural, os capítulos demonstram que a integração entre o Monitoramento Acústico Passivo (PAM) e a IA é um caminho viável e necessário para a conservação e gestão de ecossistemas marinhos. Os resultados demonstraram que as arquiteturas propostas não apenas automatizam a detecção de eventos acústicos em longas séries temporais, mas garantem uma classificação taxonômica robusta mesmo em ambientes de alta complexidade sonora.

A revisão da literatura evidenciou que algoritmos de *machine learning* (ML) e *deep learning* (DL) são fundamentais para o processamento de grandes volumes de gravações de sons de peixes. Cabe ressaltar que o sucesso dessas técnicas depende da escolha adequada do conjunto de características (*features*) e das especificidades dos sinais analisados. Nesse contexto, destacam-se limitações recorrentes na literatura, como a escassez de dados validados em nível de espécies e a necessidade de repositórios que contemplem a diversidade do repertório acústico dos peixes. Estas limitações direcionaram este trabalho a investigar não apenas a precisão dos algoritmos, mas a necessidade de métodos interpretáveis e transparentes que lidem com a escassez de dados e a complexidade das paisagens sonoras.

A aplicação de aprendizado supervisionado associada à IA explicável (XAI) demonstrou que, embora técnicas como o aumento de dados e o uso de modelos como o *Multilayer Perceptron* alcancem alto desempenho na classificação de sons pulsados de peixes em ambientes ruidosos, a confiabilidade desses sistemas está diretamente relacionada à sua interpretação. Nesse sentido, a interpretabilidade permitiu identificar quais características acústicas foram determinantes para cada classe de som. Essa etapa evidenciou que entender os mecanismos de decisão do modelo é um passo crucial para avançar além das espécies-alvo e categorizar sons de peixes desconhecidos. Destaca-se ainda, como contribuição inédita, a influência individual de cada coeficiente *Mel Frequency Cepstral Coefficient* (MFCC) para a classificação dos sinais. Enquanto pesquisas anteriores tratavam os MFCCs de forma agregada, como um conjunto único de características, este estudo demonstrou que coeficientes específicos podem atuar como descritores fundamentais com potencial para distinguir sons associados a comportamentos ou sons espécie-específicos.

De forma complementar, a utilização do modelo CLAP evidenciou o potencial dos embeddings multimodais no contexto da classificação *zero-sho*t como uma abordagem promissora diante da escassez de dados rotulados. Mesmo sem treinamento prévio no domínio bioacústico, o modelo foi capaz de capturar padrões consistentes, com destaque para a família Sciaenidae (responsável por mais de 60% dos acertos), sugerindo uma nova possibilidade para a classificação de sons produzidos por peixes. Esses resultados demonstram que a detecção e categorização de novos sinais podem ser realizadas com sucesso mesmo na ausência de treinamento prévio específico para o domínio (*zero-shot*), expandindo o alcance das ferramentas de monitoramento.

Desde a publicação da revisão da literatura apresentada no primeiro capítulo (2023), o estado da arte da aplicação de IA na bioacústica de peixes evoluiu de forma gradual, com aumento do uso do *deep learning* em comparação ao *machine learning* clássico. No entanto, essas abordagens continuam coexistindo, como observado em estudos recentes. Por exemplo, Cherubin et al. (2025) demonstraram o alto desempenho de algoritmos de *deep learning* na classificação de diferentes sonotipos de garoupas, enquanto Lancaster et al. (2025) reafirmaram a eficácia do *Random Forest* na diferenciação de sons específicos de espécies em regiões de alta diversidade. Além disso, comparações diretas (Mouy et al., 2024) indicam que, embora o DL possa superar métodos clássicos em termos de performance, sua robustez permanece sensível ao ruído ambiental e depende da disponibilidade de grandes *datasets* rotulados. Isso sugere que a escolha entre esses métodos é influenciada principalmente pelo objetivo do estudo, complexidade do cenário e disponibilidade de dados. Por fim, propostas de *active learning*, buscaram reduzir a dependência da anotação manual (Bordoux et al., 2025). Nesse contexto, as contribuições desta tese buscam preencher essas lacunas, ao incorporar a interpretabilidade e transparência na avaliação de performances dos modelos, mitigar a dependência de grandes volumes de dados rotulados e explorar a capacidade de generalização através da classificação *zero-shot* usando *embeddings* multimodais.

Em síntese, esta tese visa contribuir para o estado da arte em bioacústica de peixes marinhos, ao demonstrar que a transição de modelos supervisionados para sistemas interpretáveis e multimodais flexíveis é uma alternativa para o desenvolvimento de ferramentas com base em IA mais versáteis e com informação taxonômica. Do ponto de vista aplicado, as abordagens propostas nesta tese apresentam potencial direto para o avanço de sistemas de MAP em ambientes marinhos. Na prática, essas soluções podem ser empregadas no monitoramento contínuo de paisagens acústicas, possibilitando análises em diferentes escalas temporais (variações diárias, sazonais e interanuais) e espaciais, incluindo comparações entre habitats, regiões ou áreas sob diferentes níveis de impacto antrópico. Isso possibilitaria, por exemplo, a identificação de padrões comportamentais, bem como na identificação de alterações associadas a impactos ambientais ou pressões antrópicas. Além disso, a capacidade de reconhecer padrões acústicos mesmo na ausência de treinamento específico possibilita a aplicação em regiões de alta biodiversidade e baixo conhecimento prévio, ampliando o escopo de estudos ecológicos. Essas abordagens também têm potencial para auxiliar na gestão de áreas marinhas protegidas, por meio da avaliação da atividade biológica ao longo do tempo, contribuindo para decisões mais informadas e baseadas em dados.

Com base nos resultados e nas limitações encontradas, são sugeridos passos futuros para que esta abordagem ganhe maior robustez. Uma das sugestões seria o desenvolvimento de repositórios globais padronizados de sons de peixes, com *datasets* rotulados, facilitando o treinamento de modelos mais robustos e generalizáveis. Para o aprimoramento dos modelos multimodais, sugere-se a investigação de estratégias de otimização de *queries* (*prompt engineering*), visando identificar as descrições textuais que melhor representem as propriedades físicas e taxonômicas dos sons de peixes no espaço latente. Por fim, sugere-se o desenvolvimento de algoritmos para a integração de buscas automáticas em bases taxonômicas globais, como o *FishBase* e o *WoRMS*, por exemplo. Isso permitiria o uso de filtros biológicos e geográficos para a eliminação de candidatos improváveis durante o processo de recuperação (*retrieval*).

5.1. Referências bibliográficas

BORDOUX, V.; PARCERISAS, C.; DEBUSSCHERE, E.; JÄLMBY, M.; MURK, A. J.; VAN DER VEN, R. M. Rapid fish sound detection using human-in-the-loop active learning. *SSRN*, 2025.

CHERUBIN, L. M.; IBRAHIM, A.; SCHÄRER-UMPIERRE, M.; WOODWARD, C. Fish Acoustic Detection Algorithm Research (FADAR): a powerful tool to explore the role of call types in fish spawning behavior and trend in spawning population structure. *The Journal of the Acoustical Society of America*, v. 157, n. 4_Supplement, p. A48–A48, 2025.

LANCASTER, D.; MOUY, X.; HAGGARTY, D.; JUANES, F. Knock knock, who's there? Identifying wild species-specific fish sounds with passive acoustic localization and random forest models. *Journal of Fish Biology*, 2025.

MOUY, X.; ARCHER, S.; DOSSO, S.; DUDAS, S.; ENGLISH, P.; FOORD, C.; HAGGARTY, D. Automatic detection of fish sounds: a comparison of traditional machine learning with deep learning. *The Journal of the Acoustical Society of America*, v. 157, n. 4_Supplement, p. A49–A49, 2025.

GLOSSÁRIO

Aprendizado Ativo (*active learning*): Subárea do aprendizado de máquina em que o algoritmo é capaz de selecionar, de forma iterativa, quais dados não rotulados são mais informativos para o seu treinamento, com validação ou correção dos rótulos selecionados realizada por humanos.

Aprendizado contrastivo (*contrastive learning*): Abordagem de aprendizado de máquina que treina modelos, de modo a aproximar representações de dados semanticamente relacionados e afastar representações não relacionadas.

Aprendizado de máquina: Área da inteligência artificial dedicada ao desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e realizar tarefas como classificação, regressão e agrupamento.

Aprendizado profundo: Subárea do aprendizado de máquina baseado em redes neurais artificiais com múltiplas camadas, capaz de aprender representações hierárquicas complexas em grandes quantidades de dados.

Classificação: Tarefa de aprendizado de máquina que consiste em atribuir uma ou mais classes a uma amostra de dados.

Classificação *zero-shot*: Técnica de IA na qual modelos categorizam dados em classes para as quais não foram explicitamente treinados, explorando o conhecimento geral adquirido em modelos pré-treinados e a correspondência semântica entre a entrada e descrições em linguagem natural das classes, dispensando a necessidade de novos conjuntos de dados rotulados para cada tarefa.

***Embeddings*:** Representações vetoriais de alta dimensão que codificam informações relevantes dos dados, preservando relações de similaridade semântica e facilitando tarefas como comparação, classificação e recuperação.

***Encoder*:** Componente de uma rede neural que transforma dados brutos de entrada em uma representação matemática condensada e de alta dimensão, capturando as características mais relevantes dos dados originais.

Escala *mel*: Escala perceptual de frequência que aproxima a forma como o ouvido humano percebe variações sonoras, sendo amplamente utilizada na análise de sinais acústicos para gerar espectrogramas mais alinhados à percepção auditiva (*mel spectrograms*).

Espaço latente: Espaço matemático de dimensão reduzida onde o modelo organiza representações compactas dos dados. Nesse espaço, itens similares são mapeados em coordenadas próximas, refletindo seus padrões fundamentais em vez de sua forma bruta original.

Features: Atributos ou descritores extraídos dos dados que capturam informações relevantes para uma tarefa específica e servem como entrada para modelos de aprendizado de máquina.

Frequência *mel*: Representação da frequência convertida da escala linear (em Hertz) para a escala mel, enfatizando diferenças perceptualmente mais relevantes em determinadas faixas de frequência.

Modelos multimodais: Modelos computacionais capazes de processar e integrar informações provenientes de diferentes modalidades de dados, como texto, imagem e áudio, aprendendo representações conjuntas.

Pipeline: Estrutura organizada de etapas sequenciais de processamento que define o fluxo de dados desde a entrada até a obtenção dos resultados finais.

Query: Entrada fornecida a um sistema ou modelo com o objetivo de realizar uma consulta, busca ou inferência, servindo como referência para comparação ou recuperação de informações.

Redes neurais: Modelos computacionais inspirados no cérebro humano, compostos por unidades interconectadas que aprendem relações complexas a partir de dados.

Redes neurais convolucionais: Arquitetura específica de redes neurais projetada para aprender padrões locais e hierárquicos em dados estruturados em grade, por meio de operações de convolução.

Tarefa *downstream*: Aplicação ou problema específico que se beneficia de modelos ou representações aprendidas em uma etapa prévia e mais geral de pré-treinamento (*upstream*), na qual um modelo pré-treinado é adaptado para o objetivo final.

Apêndice 1 - SUPPLEMENTARY MATERIAL TO THE PAPER "OPTIMAL FEATURE SELECTION AND MODEL EXPLANATION FOR REEF FISH SOUND CLASSIFICATION" (CAPÍTULO 2)

Tables

TABELA S1 - Feature sets used to classify reef fish sounds. These features form sets 1 and 2. Feature set 3 consists of all the features, extracted from sets 1 and 2.

feature set	feature	domain	description
1	call duration	time	the total duration in milliseconds (ms) of a fish call
	pulse duration	time	the duration in milliseconds (ms) of a single pulse within a fish call
	interpulse interval (IPI)	time	the interval in milliseconds between 2 consecutive pulses within a fish call
	number of pulses	time	the total number of pulses within a fish call
	low frequency	frequency	the lower frequency bound of the selection[1]
	high frequency	frequency	the upper frequency bound of the selection[1]
	energy	amplitude	the total energy within the selection bounds[1]
	peak frequency	frequency	the frequency range in which the peak power occured[1]
	center frequency	frequency	the smallest discrete frequency within the selection[1]
	tempo	time	estimates the tempo (beats per minute)[2]
2			

	mel frequency cepstral coefficients (MFCCs)	cepstral	represent the general spectral envelope of signals for each frame (CHARIF et al., 2010).
	spectral centroid	frequency	computes the spectral centroid by normalising each frame of a magnitude spectrogram and treating it as a distribution over frequency bins from which the mean (centroid) per frame is extracted.[2]
	spectral rolloff	frequency	is the centre frequency for each frame for a spectrogram bin such that at least roll percent (0.85 by default) of the energy of the spectrum in that frame is contained in that bin and the bins below it.
	spectral flux	frequency	computed as the squared difference between the normalized magnitudes of the spectra of two successive frames, it measures the spectral change between them (ZELKOWITZ, 2010).
	zero crossing rate (ZCR)	frequency	measures the number of times the waveform crosses the zero axis (GIANNAKOPOULOS & PIKRAKIS, 2014).

[1] These features are measurements extracted from spectrograms using Raven software. Their descriptions were extracted from Raven User's Manual (CHARIF et al., 2010).

[2] The descriptions of these features were extracted from Librosa's documentation (MCFEE et al., 2015).

TABELA S2 - Performance of supervised classification for other tested feature sets (original and augmented). These sets consist of (1) MFCCs only, (2) spectral features from the main feature sets 1 and 2, and (3) temporal features from the main feature sets 1 and 2. (4) TTFs (spectral centroid, spectral rolloff, spectral flux, zero-crossing rate The performance metrics are Ac (% accuracy), Pc (% precision), Rc (% recall) and f1 (% f1-score).

classifier	**feature set**	**original**				**augmented**			
		Ac	**Pc**	**Rc**	**f1**	**Ac**	**Pc**	**Rc**	**f1**
	MFCCs only	77.8	73.7	73.2	73.1	83.3	83.4	82.2	82.4
	spectral	72.2	72.3	71.7	69.3	75.0	77.9	75.0	75.3
	temporal	55.6	57.8	56.5	54.7	75.0	75.1	73.8	73.8
NB	**TTFs**	80.6	77.8	73.4	76.7	85.2	84.9	84.3	84.5
RF	**MFCCs only**	72.2	70.8	69.6	69.5	88.9	88.6	88.2	88.2

	spectral	80.6	79.7	79.8	77.6	84.2	84.5	83.7	83.6
	temporal	52.8	55.9	53.4	52.2	76.8	77.0	75.9	76.3
	TTFs	69.4	68.6	67.9	67.9	88.9	88.3	88.2	88.2
	MFCCs only	69.4	67.2	67.9	67.2	74.1	73.5	73.1	73.2
	spectral	72.2	75.0	73.1	71.1	83.3	83.7	82.9	82.9
	temporal	52.8	54.3	53.4	52.2	75.0	74.5	74.6	74.4
DT	**TTFs**	66.7	64.7	64.7	64.2	77.8	77.4	77.0	77.1
	MFCCs only	66.7	73.4	68.0	66.9	88.0	88.9	87.4	87.5
	spectral	69.4	65.7	66.6	66.1	89.8	89.9	89.6	89.6
	temporal	75.0	72.5	71.3	71.6	73.1	74.4	73.7	73.3
MLP	**TTFs**	72.2	77.7	73.0	71.5	91.7	92.1	91.7	91.6

TABELA S3 - Comparison of the performance of the classifiers with average mean and standard deviation after a 5-fold cross-validation for the original and the augmented dataset. The performance metrics are Ac (% accuracy), Pc (% precision), Rc (% recall) and f1 (% f1-score).

classifier	**feature set**	**original**				**augmented**			
		Ac	**Pc**	**Rc**	**f1**	**Ac**	**Pc**	**Rc**	**f1**
	1	84∓0.8	90∓0.2	84∓0.8	83∓0.9	87∓0.8	88∓0.8	87∓0.8	87∓0.8
	2	72∓2.4	79∓1.7	72∓2.4	72∓2.5	81∓0.8	83∓0.5	81∓0.8	81∓0.8
NB	**3**	88∓0.8	91∓0.7	88∓0.8	88∓0.8	94∓0.5	94∓0.4	94∓0.5	94∓0.4
	1	85∓1.5	88∓0.6	82∓1.7	84∓1.8	94∓0.8	95∓0.4	94∓0.7	94∓0.5
	2	75∓1.3	75∓1.3	72∓0.8	73∓0.7	89∓0.5	88∓0.6	89∓0.6	86∓0.3
RF	**3**	89∓1.2	91∓0.8	88∓1.2	85∓1.7	93∓0.6	94∓0.4	94∓0.4	95∓0.4
	1	81∓1.5	85∓0.9	81∓1.5	80∓0.2	92∓0.5	93∓0.6	90∓0.6	92∓0.6
	2	68∓1.9	71∓2.1	68∓1.9	66∓1.8	73∓1.5	76∓0.8	69∓1.6	71∓1.5
DT	**3**	81∓2.0	85∓1.9	81∓2.4	78∓2.1	88∓0.9	91∓0.6	90∓0.6	92∓0.6
	1	88∓1.1	91∓0.9	88∓1.1	88∓1.1	95∓0.9	95∓0.8	95∓0.9	95∓0.9
	2	74∓2.1	80∓2.3	73∓2.1	73∓2.3	87∓0.9	88∓0.8	87∓0.8	87∓0.9
MLP	**3**	89∓0.9	90∓1.0	89∓0.8	89∓0.9	95∓0.6	95∓0.6	95∓0.6	95∓0.6

TABELA S4 - Performance metrics of RF classifier by sound class for each feature set. The performance metrics are Ac (% accuracy), Pc (% precision), Rc (% recall) and f1 (% f1-score).

feature set	class	performance			
		Ac	Pc	Rc	f1
	1	94	91	83	87
	2	93	92	92	92
	3	93	92	100	96
1	4	93	97	97	97
	1	87	70	84	76
	2	87	89	75	81
	3	86	96	93	95
2	4	87	90	97	93
	1	93	87	89	89
	2	92	89	91	91
	3	92	93	93	93
3	4	92	100	94	97

Figures

FIGURA S1 - Summary of the SHAP values for each class for feature set 3. (a) Class 1. (b) Class 2. (c) Class 3. (d) Class 4. The features are plotted on the y-axis, the x-axis indicates higher or lower values for the influence on predictions, and the colour indicates high or low values. (The plot for Class 1 is in the main document).

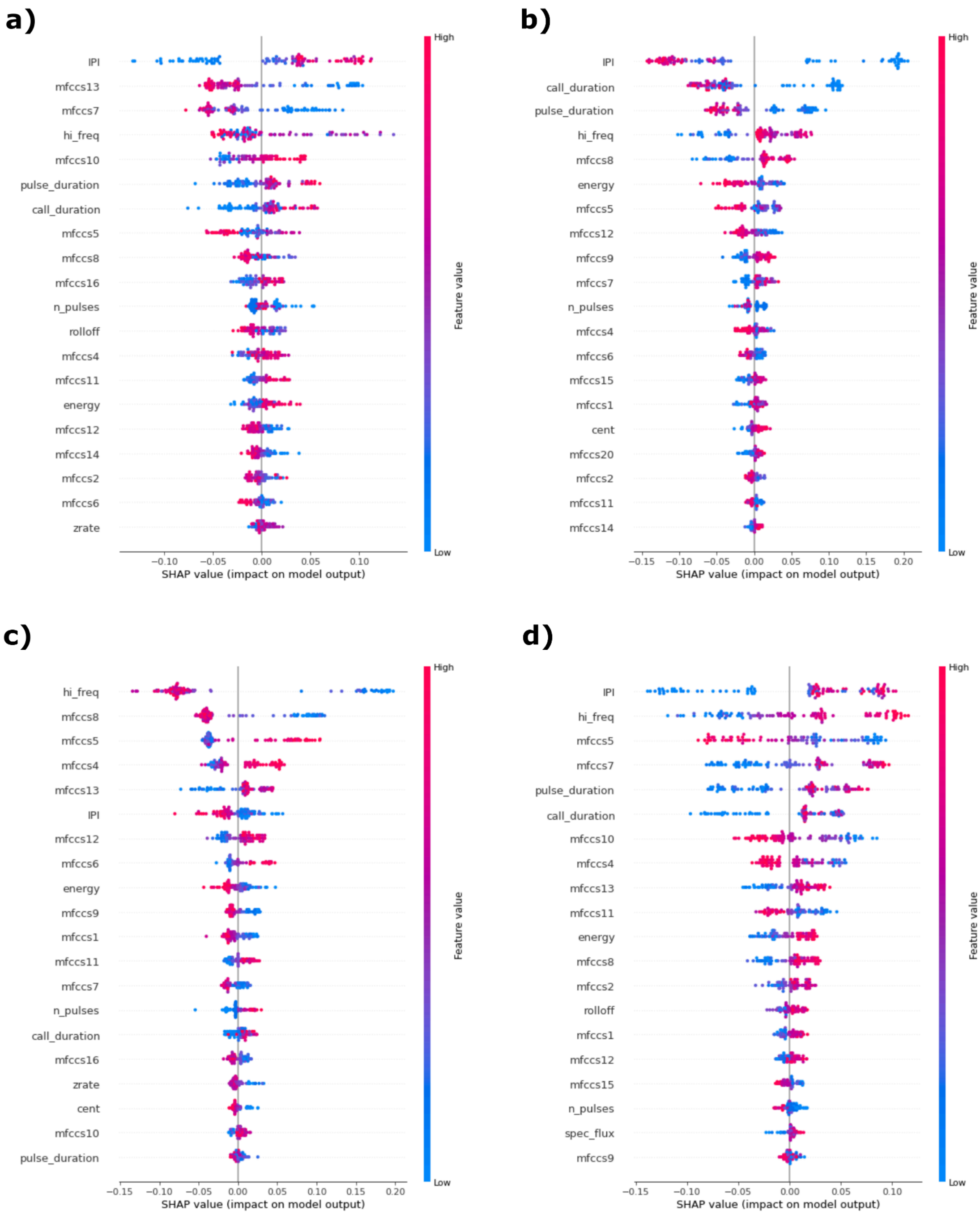


FIGURA S2 - Summary of the SHAP values for each class for feature set 1. (a) Class 1. (b) Class 2. (c) Class 3. (d) Class (4). The features are plotted on the y-axis, the x-axis indicates higher or lower values for the influence on predictions, and the colour indicates high or low values.

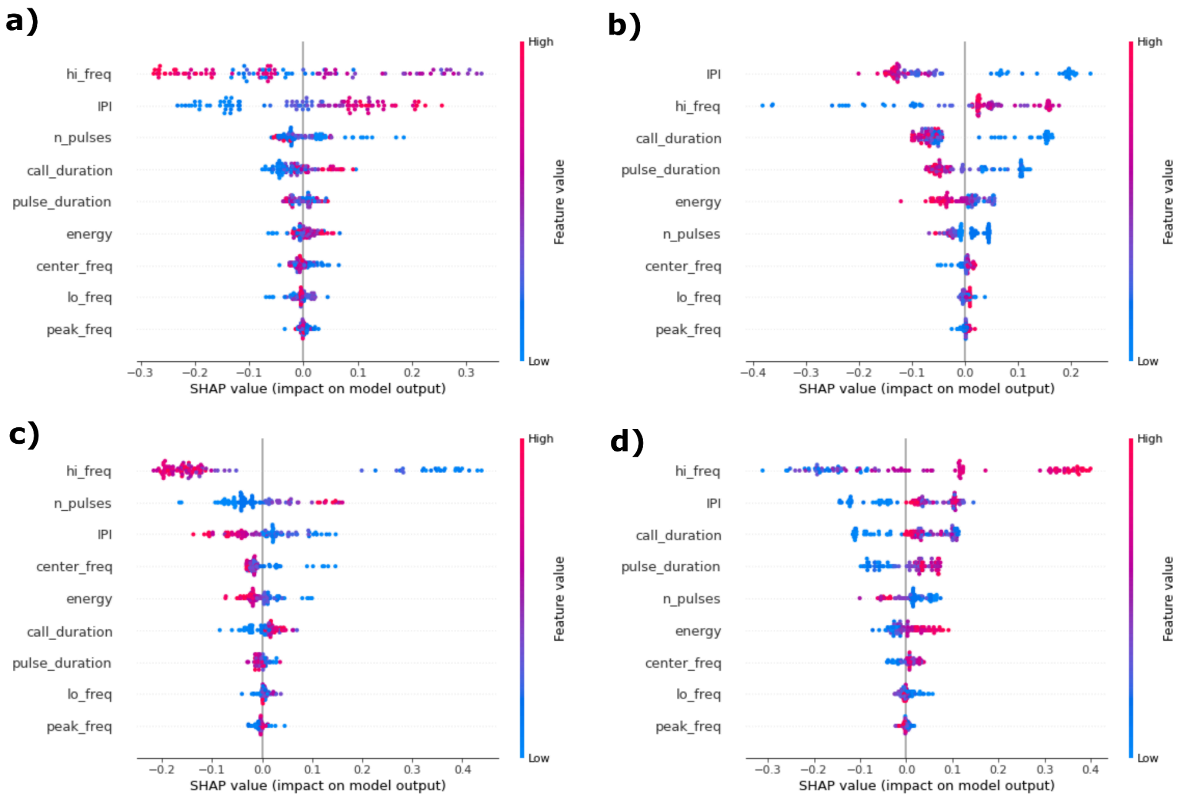


FIGURA S3 - Summary of the SHAP values for each class for feature set 2. (a) Class 1. (b) Class 2. (c) Class 3. (d) Class (4). The features are plotted on the y-axis, the x-axis indicates higher or lower values for the influence on predictions, and the colour indicates high or low values.

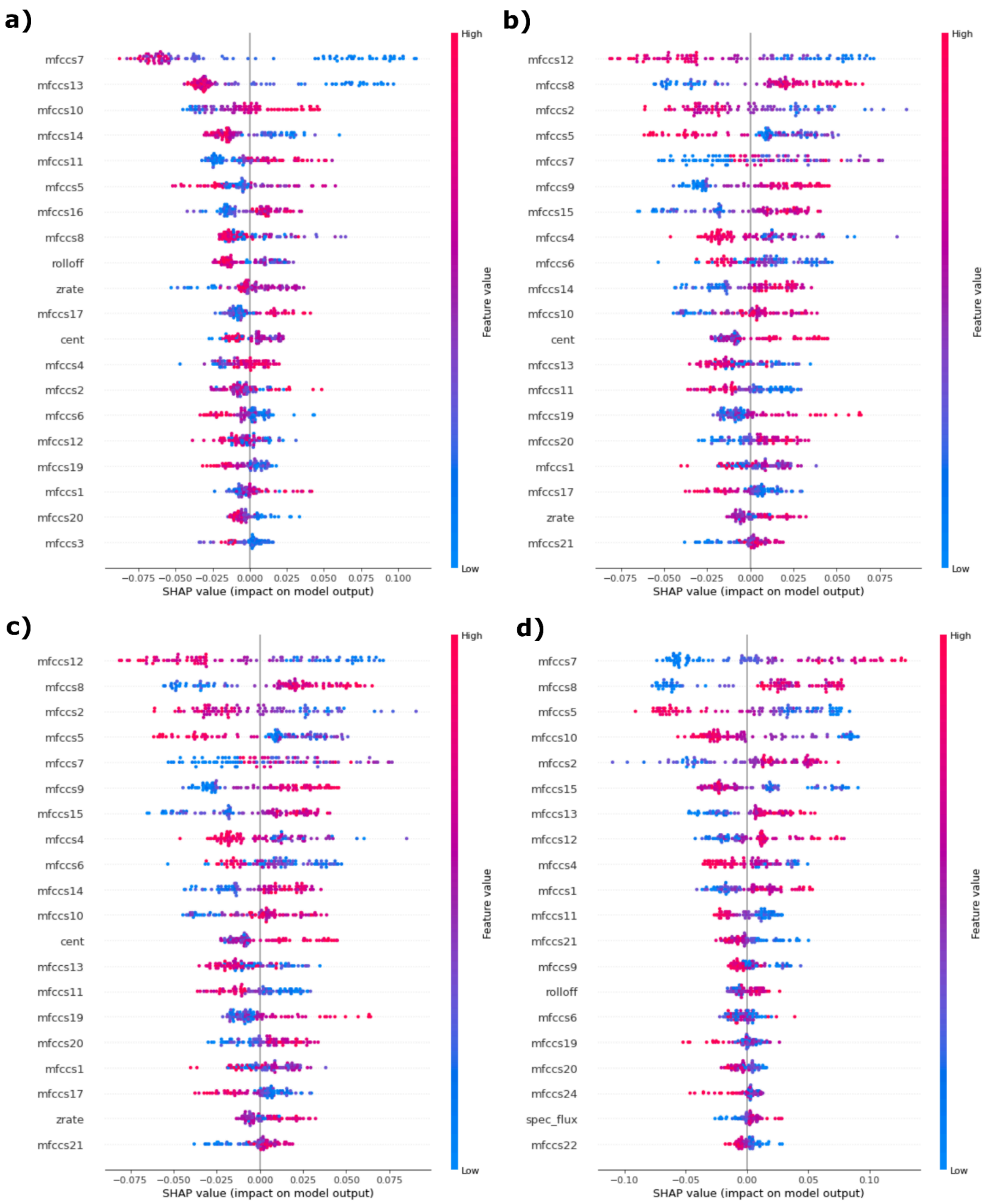


FIGURA S4 - Summary of the influence of acoustic features in predicting each class of fish sounds for the TTFs set. This set consists of MFCCs, spectral centroid, spectral rolloff, ZCR, and spectral flux.

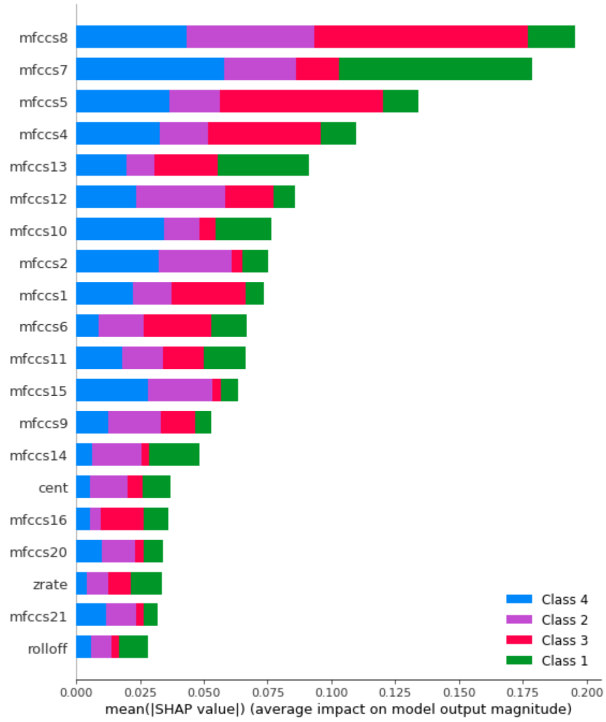


FIGURA S5 - Summary of the influence of acoustic features in predicting each class of fish sounds for the spectral feature set. This set is the combination of all spectral features, including TTFs.

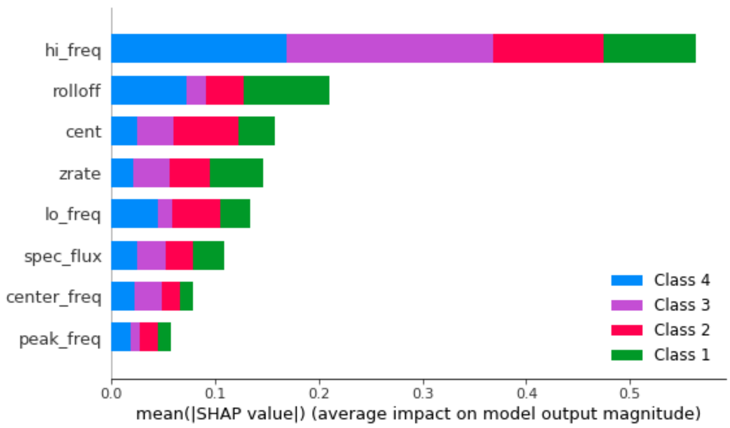


FIGURA S6 - Dependence plot for mfcc8 and high frequency for Class 3 in the feature set 3. The plot shows the influence of high frequency on the vertical distribution of SHAP values for MFCC5.

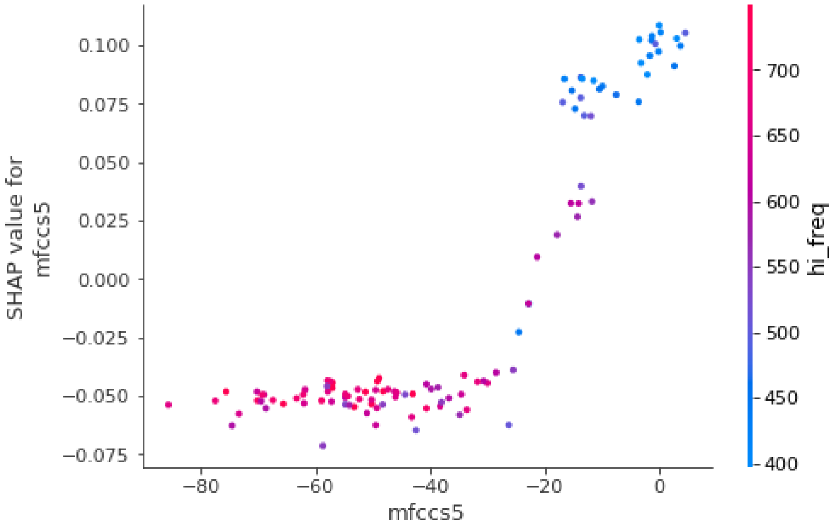


FIGURA S7 - Dependence plot for IPI and call duration for Class 2 in the feature set 3. The plot shows the influence of call duration on the vertical distribution of SHAP values for IPI.

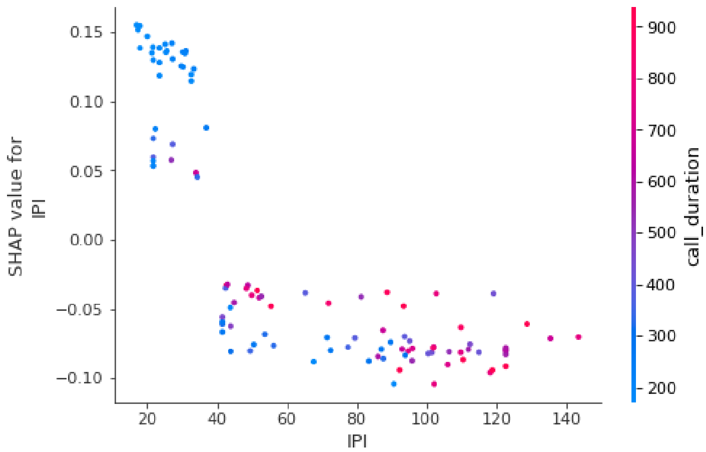


FIGURA S8 - Summary of the influence of acoustic features in predicting each class of fish sounds for Feature set 1. This set consists of energy, spectral and temporal features extracted with Raven Pro 1.6.

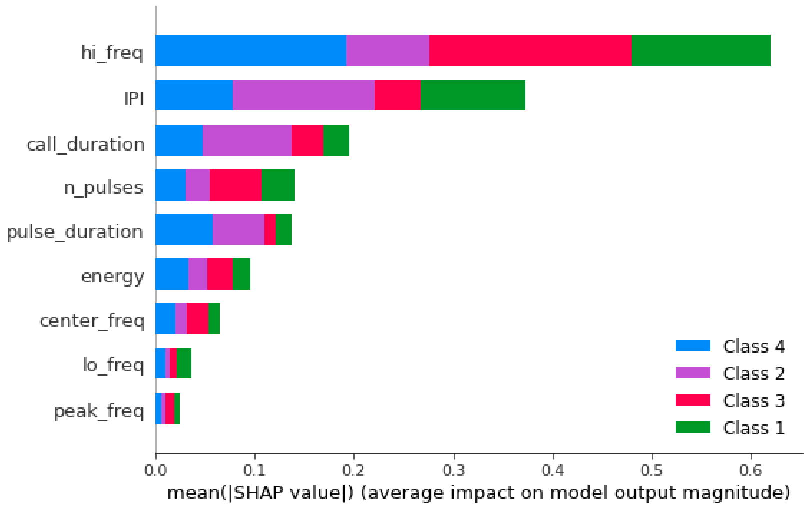
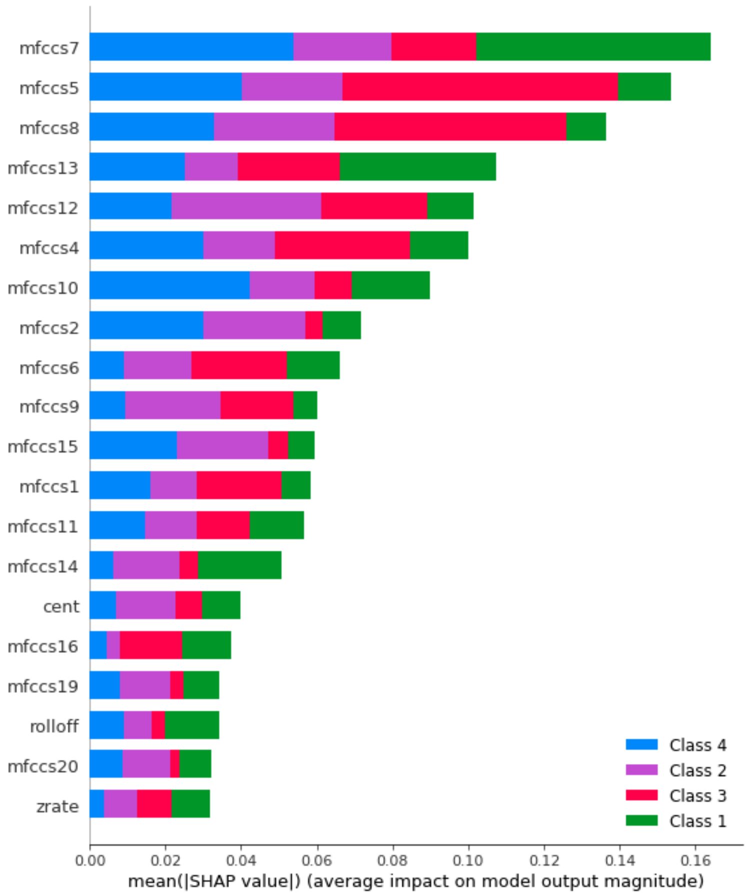


FIGURA S9 - Summary of the influence of acoustic features in predicting each class of fish sounds for Feature set 2. This set consists of tempo, spectral and temporal features extracted with Librosa library.



Methods

The main characteristics of the local soundscape

Cabo Frio island is characterised by rocky shores with a high coverage of invertebrates, such as shrimps, bivalves, barnacles, and sea urchins, and a high diversity of fish. This is a tourist area with intensive recreational boating at certain periods of the year. The local soundscape is characterised by remarkable diel variation caused mainly by choruses of invertebrates and fishes. Fish choruses produce high average peak power in the low frequency band (< 2 kHz) during dawn and dusk, while snapping shrimp sounds often dominate the twilight period. Recreational boats and fishing vessels also contribute to this variation, as they operate more intensively during the day. During the high season, the soundscape has a significant contribution of recreational boats (JESUS et al., 2020; CAMPBELL et al., 2019).

Fish sound classes details

In order to determine whether an acoustic signal within a recording was a fish sound, we analysed the sound files both aurally and visually, inspecting spectrograms and oscillograms using Raven Pro software. We then compared the putative fish sounds with the recordings of fish sounds available in sound libraries, such as Fish Sounds (LOOBY, 2023), Fish Base (FROESE; PAULY, 2021), and the GSO Fish Sounds Collection (FISH; MOWBRAY, 1970). Besides, the frequency range, and temporal patterns, such as duration and number of pulses of the four classes are consistent with the literature describing fish sounds (AMORIM, 2006; LADICH, 2014; CARRIÇO et al., 2019; DESIDERÀ et al., 2019). Thus, we only selected clearly identifiable as fish sounds. We chose the classes based on the frequency of occurrence in the dataset. These were the four most frequent sound types. We verified that they were distinct sound classes by assessing the temporal and spectral parameters of these pulsed sounds, which could be discriminated both aurally and visually. After identifying the target sounds in the wav files, we generated spectrograms in Raven Pro software and used the selection tool to delineate the region of interest of the fish acoustic signals in both the waveform and spectrograms.

Details of the data augmentation process

The augmentation for each dataset was performed using the standard deviation of the feature values as a guide. We computed the standard deviation of each feature and created new samples by adding and subtracting the standard deviation to each feature value. Thus, for each original sample, two new synthetic samples were created (one by adding and one by subtracting the standard deviation from the original sample). We made this process for each class in each dataset, which originally had 120 samples. By doing this, we reached a total of 360 samples (90 per class).

Details of the model explanation

After training the RF classifier on the augmented datasets, we used the 'TreeExplainer' from the SHAP library (LUNDBERG; LEE, 2017), which is optimised for tree-based models, such as random forests. Then, SHAP values for each feature in the dataset were computed, quantifying the contribution of each feature for individual predictions of fish sound classes. SHAP values were then computed for each feature in the dataset, quantifying the contribution of each feature to individual predictions of fish sound classes. Finally, feature importance, SHAP summary and dependence plots were generated.

SHAP is a technique that explains individual predictions by additive feature attribution (MOLNAR, 2019). It assigns an importance value to each feature, explaining how a particular feature contributes to moving the prediction towards or away from a particular class. The SHAP framework can be summarised as follows: (1) it starts with a baseline, which is the average of all predictions; (2) it produces attributions for each feature; (3) it assesses the impact on the model prediction when a particular feature is included; (4) for each feature, it calculates the average over all possible feature combinations and assesses its positive or negative contributions to the predictions (LUNDBERG; LEE, 2017; MOLNAR, 2019; SUNDARARAJAN; NAJMI, 2020).

Apêndice 2 - SUPPLEMENTARY MATERIAL TO THE PAPER "EXPLORING THE FEASIBILITY OF MULTIMODAL EMBEDDINGS FOR FISH SOUND CLASSIFICATION" (CAPÍTULO 3)

Tables

TABELA S1 - Distribution of sonotypes, species, and genera within each family in the dataset.

Dataset composition			
Family	**Genus**	**Species**	**Sonotypes**
Acanthuridae	1	3	11
Albulidae	1	1	3
Alosidae	2	2	2
Anguillidae	1	1	1
Anoplopomatidae	1	1	4
Ariidae	1	1	2
Balistidae	4	6	16
Batrachoididae	4	5	28
Carangidae	8	15	21
Carapidae	1	3	3
Centropomidae	1	1	1
Chaetodontidae	1	2	2
Clupeidae	1	2	4
Cottidae	1	3	10
Dactylopteridae	1	1	13
Diodontidae	2	3	13
Dorosomatidae	1	1	1
Ephippidae	1	1	3
Epinephelidae	5	13	34
Gadidae	4	5	26
Gerreidae	3	4	6
Glaucosomatidae	1	1	2
Gobiidae	2	2	4

Grammistidae	1	1	1
Haemulidae	5	14	30
Holocentridae	2	2	6
Kyphosidae	1	1	6
Labridae	5	8	9
Lotidae	1	1	2
Lutjanidae	3	7	15
Megalopidae	1	1	1
Monacanthidae	4	4	9
Moronidae	1	1	1
Mullidae	2	2	2
Ophidiidae	1	2	7
Ostraciidae	1	4	9
Pempheridae	1	1	1
Phycidae	1	2	6
Pomacanthidae	2	5	16
Pomacentridae	5	8	11
Pomatomidae	1	1	1
Salmonidae	3	4	9
Scaridae	1	5	9
Sciaenidae	15	23	54
Serranidae	3	3	4
Sparidae	6	7	19
Sphyraenidae	1	1	1
Syngnathidae	1	1	1
Tetraodontidae	2	3	13
Triglidae	3	5	21

Apêndice 3 - Publicações

Artigos publicados

BARROSO, V. R.; LESSA, A. A.; FERREIRA, C. E.; XAVIER, F. C. Optimal feature selection and model explanation for reef fish sound classification. *Philosophical Transactions B*, v. 380, n. 1928, p. 20240055, 2025.

LESSA, A. A.; XAVIER, F. C.; **BARROSO, V. R**.; CORDEIRO, C. A.; FERREIRA, C. E. Exposure to anthropogenic noise affects feeding but not territory defence in damselfishes. *Animal Behaviour*, p. 123130, 2025.

LUZ, A. C. N.; **BARROSO, V. R.**; BATISTA, D.; LESSA, A. A.; COUTINHO, R.; XAVIER, F. C. Early detection of marine bioinvasion by sun corals using YOLOv8. *Intelligent Marine Technology and Systems*, v. 3, n. 1, p. 2, 2025.

BARROSO, V. R.; XAVIER, F. C.; FERREIRA, C. E. L. Applications of machine learning to identify and characterize the sounds produced by fish. *ICES Journal of Marine Science*, v. 80, n. 7, p. 1854–1867, 2023.

Capítulos de livros publicados

LESSA, A. A.; **BARROSO, V. R.**; XAVIER, F. C.; FERREIRA, C. E. Impacts of anthropogenic sounds on reef fish. In: *The effects of noise on aquatic life: principles and practical considerations*. Cham: Springer International Publishing, 2024. p. 877–885.

MINELLO, M.; **BARROSO, V. R.**; LESSA, A. A.; HOFFMANN, L. S.; ARAÚJO, S. A.; RAMOS, Y. F.; FAGUNDES NETTO, E. B.; XAVIER, F. C.; PARO, A. D.; MELO JÚNIOR, U. G. A acústica submarina como ferramenta de monitoramento ambiental. In: SOUTO, R. D. (ed.). *Gestão ambiental e sustentabilidade em áreas costeiras e marinhas: conceitos e práticas*. v. 2. [S.l.]: Raquel Dezidério Souto, 2022. p. 433–461.

Apêndice 4 - Atividades de revisões por pares em periódicos científicos

- *Artificial Intelligence Review* - Revisora (Agosto 2025)
- *Artificial Intelligence Review* - Revisora (Setembro 2025)
- *Ecological Informatics* - Revisora (Agosto 2025)
- *Global Ecology and Conservation* - Revisora (Setembro 2025)
- *Global Ecology and Conservation* - Revisora (Novembro 2025)