Deep Learning Based Frameworks for the Detection and Classification of Soniferous Fish

Ziqi Huang,[1,2,a)] Dominik Ochs,[3,4,a)] M. Clara P. Amorim 0000-0002-2453-6999,[5,6] Paulo J. Fonseca 0000-0002-2663-9385,[7,8] Mayank Goel 0000-0003-1237-7545,[9] Nuno Jardim Nunes 0000-0002-2498-0643,[1,2] Manuel Vieira 0000-0002-3103-8330,[5,b)] and Manuel Lopes 0000-0002-6238-8974[3,2,b)]

[1)] *ITI / LARSyS, Lisbon, Portugal*

[2)] *Instituto Superior Técnico, Universidade de Lisboa, Portugal*

[3)] *INESC-ID, Lisbon, Portugal*

[4)] *Heidelberg University, Germany*

[5)] *MARE – Marine and Environmental Sciences Centre / ARNET - Aquatic Research Network, Universidade de Lisboa, Lisboa, Portugal*

[6)] *Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisboa, Portugal*

[7)] *cE3c - Center for Ecology, Evolution and Environmental Changes & CHANGE - Global Change and Sustainability Institute, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal*

[8)] *Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*

[9)] *Carnegie Mellon University, USA*

Passive Acoustic Monitoring (PAM) is emerging as a valuable tool for assessing fish populations in natural habitats. This study compares two deep learning-based frameworks: (1) a multi-label classification system (SegClas) combining Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) networks and, (2) an object detection approach (ObjDet) using a YOLO-based model to detect, classify, and count sounds produced by soniferous fish in the Tagus estuary, Portugal. The target species-Lusitanian toadfish (*Halobatrachus didactylus*), meagre (*Argyrosomus regius*), and weakfish (*Cynoscion regalis*)-exhibit overlapping vocalization patterns, posing classification challenges. Results show both methods achieve high accuracy (over 96%) and F1 scores above 87% for species-level sound identification, demonstrating their effectiveness under varied noise conditions. ObjDet generally offers slightly higher classification performance (F1 up to 92%) and can annotate each vocalization for more precise counting. However, it requires bounding-box annotations and higher computational costs (inference time of ca. 1.95 seconds per hour of recording). In contrast, SegClas relies on segment-level labels and provides faster inference (ca. 1.46 seconds per hour). This study also compares both counting strategies, each offering distinct advantages for different ecological and operational needs. Our results highlight the potential of deep learning-based PAM for fish population assessment.

[https://doi.org(DOI number)]

[XYZ] Pages: 1–13

I. INTRODUCTION

Monitoring marine ecosystems is crucial for protecting biodiversity and maintaining ecological balance (Watson *et al.*, 2019). Traditional survey methods, such as trawling and visual observations, can be costly, labour-intensive, and often disruptive to aquatic life, or even impossible in certain locations/depths. In contrast, passive acoustic monitoring (PAM) is emerging as an attractive alternative for continuous, non-intrusive assessment of underwater soundscapes, enabling researchers to capture the presence and behaviour of marine organisms through their vocalizations (Boelman *et al.*, 2007; Kvsn *et al.*, 2020; Ribeiro *et al.*, 2022). Ecoacoustic data from marine soniferous animals can provide insights into reproduction, niche disputes, distribution and potential habitat shifts — all critical information for ecological management and conservation strategies (Amorim *et al.*, 2023; Bolgan *et al.*, 2023; Marques *et al.*, 2013; Stratoudakis *et al.*, 2024; Van Hoeck *et al.*, 2021).

The Tagus Estuary in Portugal presents an ideal setting to advance these monitoring approaches, given its ecological complexity and the co-occurrence of multiple highly soniferous fish species, including the Lusitanian

[a)] Equal contribution to the algorithm development and data analysis.

[b)] Equal contribution to the scientific coordination toadfish (*Halobatrachus didactylus*), meagre (*Argyrosomus regius*), and weakfish (*Cynoscion regalis*) (Amorim *et al.*, 2023; Vieira *et al.*, 2021a). These species often produce overlapping calls, further complicated by confounding factors such as significant intra-specific variation on the toadfish's vocal repertoire, minimal inter-specific variation between the meagre and weakfish's calls, variable ambient noise (both natural and anthropogenic), and varying distances between fish and hydrophones (Amorim *et al.*, 2008, 2023; Vieira *et al.*, 2021a). Traditional acoustic detection systems relying solely on amplitude thresholds or human-driven annotation can struggle with such complex acoustic scenes, especially when signal-to-noise ratios are low or multiple species vocalize simultaneously (Guyot *et al.*, 2021).

Recent advances in deep learning have provided powerful tools for analysing high-dimensional signals and extracting robust features directly from data (Mouy *et al.*, 2024). Among these, convolutional neural networks (CNNs) are particularly effective in identifying localized spectral structures (Rippel *et al.*, 2015). In contrast, recurrent architectures such as long short-term memory (LSTM) networks excel at capturing temporal dependencies (Lai *et al.*, 2018). Meanwhile, object detection frameworks exemplified by You Only Look Once (YOLO)-based models offer a complementary strategy by annotating individual call instances in time-frequency representations through a single forward pass (Jiang *et al.*, 2022). Due to its relatively small model size and high inference speed, both approaches can directly output class predictions, especially suited for real-time tasks in complex underwater acoustic scenes. In aquatic bioacoustics, these methods help address challenges posed by overlapping vocalizations, unpredictable noise conditions, and class imbalance. This offers an advantage over other developed systems for recognizing fish sounds (Malfante *et al.*, 2018; Monczak *et al.*, 2019; Vieira *et al.*, 2015) , which often struggle with overlapping vocalizations, subtle sound type differences (e.g., meagre vs. weakfish, as noted by Amorim *et al.* (2023)), and slow inference speeds. Moreover, the use of CNNs is supported by widely available deep learning libraries, making these tools more accessible to researchers without a computational background.

In this study, we propose and compare two deep learning approaches for multi-label classification and counting of fish vocalizations in the Tagus estuary. The first, a multi-label segmentation-based classification system (SegClas), segments the audio into fixed intervals and uses a hybrid CNN–LSTM model to capture spectral and temporal features of each segment. The second, an object detection approach (ObjDet), employs a YOLO-based framework to detect and localize calls within spectrograms, thus enabling a more fine-grained count of individual vocalizations. Both methods integrate data augmentation strategies to address noise variability and aim to provide scalable solutions for real-world monitoring scenarios. We evaluate these systems on multiple metrics to assess their capacity for long-term, fully automated fish monitoring.

II. METHODS

A. Data Description

The full dataset comprises 8.5 years of continuous recordings from the Tagus estuary, Portugal (April 22, 2016–August 15, 2024), collected using a High Tech 94 SSQ hydrophone (sensitivity of $-165dB$ re $1V/\mu Pa$; $+/-1dB$ from from 30Hz to 6 kHz; High Tech Inc., Gulfport, MS, USA) anchored ca. 20 cm above the bottom. Data were recorded by a 16-bit, 16-channel logger (Measurement Computing Corporation LGR-5325) at 22 kHz in 2016 (later down-sampled to 4 kHz) and at 4 kHz from 2017 onward. Depth at the site varied with tide (2–6 m). Due to logistic reasons, recordings are missing for the periods Oct 2019–Feb 2020 and July–Nov 2023. Tagus Estuary is an environment shared by three target fish species: Lusitanian toadfish, meagre, and weakfish, with the latter two sharing similar ecological and acoustic niches (Amorim *et al.*, 2023). The sounds made by all three species overlap considerably in both temporal and frequency domains, leading to notable classification challenges(Amorim *et al.*, 2023; Vieira *et al.*, 2021a). While sounds produced by meagre and weakfish are typically detected between 100 and 800 Hz, and those of toadfish between 50 and 600 Hz, all three species can produce sounds up to 1 kHz. Temporally, meagre grunt calls typically consist of up to approximately 100 pulses, with a pulse period generally between 16 and 22 ms. In contrast, weakfish grunt calls comprise 3 to 14 pulses, with a pulse period ranging from 50 to 90 ms (Amorim *et al.*, 2023). Notably, toadfish boatwhistles can resemble meagre calls, and toadfish grunt trains can approximate weakfish vocalizations in both temporal and frequency characteristics when the signal-to-noise ratio is low, further complicating species-level automatic classification. See supplementary materials for a visuallization of the different patterns. Additionally, frequent passage of small boats and local ferries near the logger introduces further challenges to automation (Vieira *et al.*, 2020, 2021b). Informed by this expert knowledge of these species' vocalizations, recordings for the training and test datasets were manually annotated using Raven Pro 1.6.5[1], through combined visual inspection of spectrograms and aural validation (see detail of each species' calls in Fig.S1). A multi-label annotation scheme was employed to indicate species presence without differentiating call types. Subsequently, audio was segmented into 3-second clips — a duration selected to balance the capture of complete vocalizations with the need to minimize overlapping sounds or noise.

Figure 1 summarizes the distribution of recordings in the training and test datasets. The core training data comes from six days in July 2021, representing a constrained scenario with limited data. To improve generalization across time and acoustic conditions, the train-

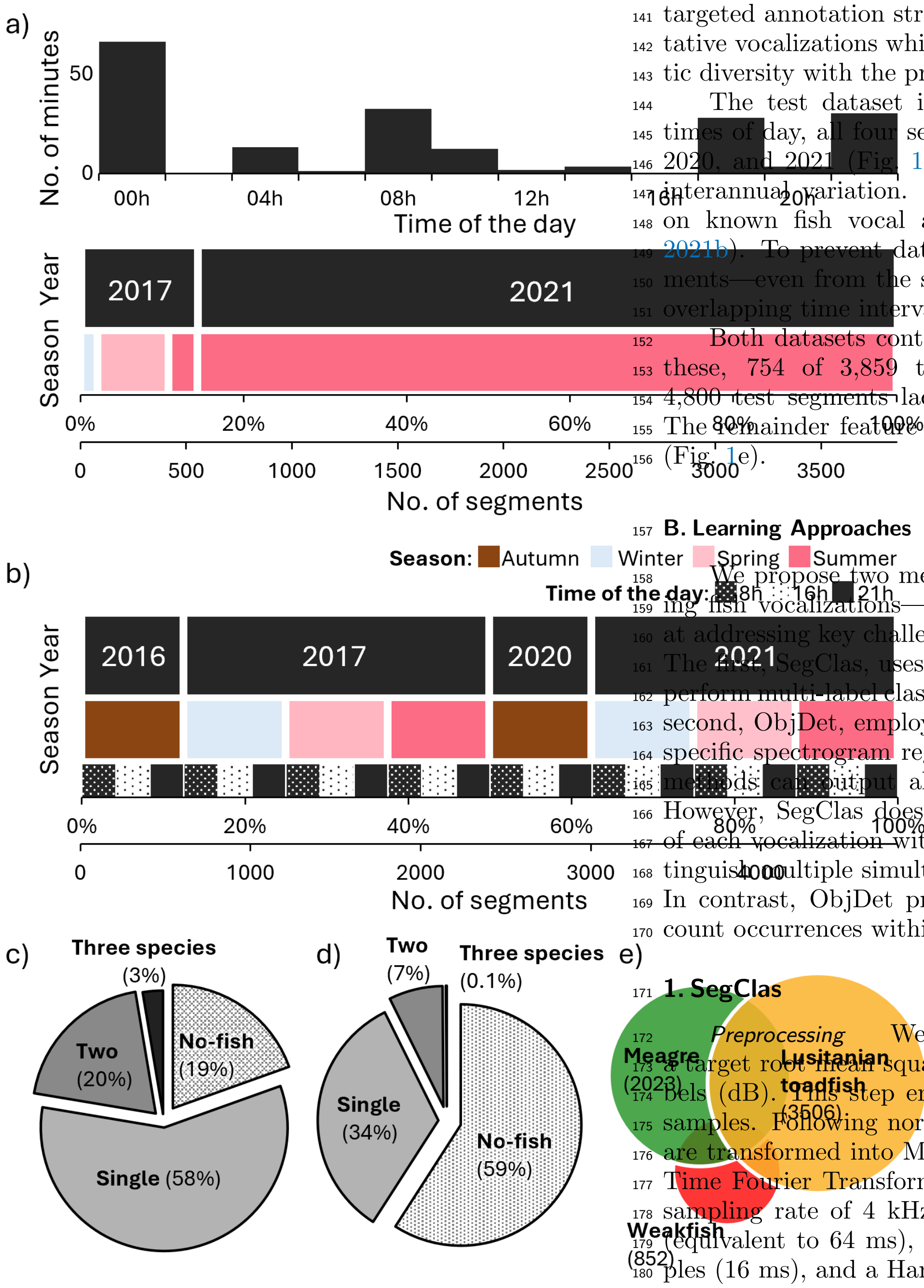


FIG. 1. Training and test data distribution. (a) Characterization of the training dataset, showing temporal distribution by year, season and time of day. (b) Representation of the test data across four years, all seasons, and three times of day. (c-d) Proportional breakdown of segments by label combinations in the (c) training and (d) test datasets, using a multi-label one-hot encoding scheme: $Y = (y_t, y_m, y_w)$ for toadfish, meagre, and weakfish, respectively. (e) Diagram showing the distribution of labeled data for Lusitanian toadfish, meagre, and weakfish in both datasets.

ing set was supplemented with short annotated segments from three additional days in 2017: 4 minutes from January, 17 from April, and 7 from July (Fig. 1a). This targeted annotation strategy aimed to capture representative vocalizations while balancing temporal and acoustic diversity with the practical limits of manual labeling.

The test dataset includes recordings from various times of day, all four seasons, and the years 2016, 2017, 2020, and 2021 (Fig. 1b), capturing diel, seasonal, and interannual variation. Time points were chosen based on known fish vocal activity patterns (Vieira *et al.*, 2021b). To prevent data leakage, training and test segments—even from the same day—were taken from non-overlapping time intervals.

Both datasets contain fish and non-fish sounds. Of these, 754 of 3,859 training segments and 2,837 of 4,800 test segments lack fish vocalizations (Fig. 1c,d). The remainder feature single- or multi-species fish calls (Fig. 1e).

B. Learning Approaches

We propose two methods for detecting and classifying fish vocalizations—summarized in Figure 2—aimed at addressing key challenges in fish vocalization analysis. The first, SegClas, uses a CNN followed by an LSTM to perform multi-label classification on audio segments. The second, ObjDet, employs YOLO to identify and classify specific spectrogram regions for each fish species. Both methods can output all potential labels in a segment. However, SegClas does not pinpoint the exact location of each vocalization within the segment, nor does it distinguish multiple simultaneous calls of the same species. In contrast, ObjDet provides bounding boxes and can count occurrences within a segment.

1. SegClas

Preprocessing We begin by normalizing each clip to a target root mean square (RMS) level specified in decibels (dB). This step ensures consistent loudness across samples. Following normalization, the processed signals are transformed into Mel spectrograms using the Short-Time Fourier Transform (STFT). Specifically, we use a sampling rate of 4 kHz, a window size of 256 samples (equivalent to 64 ms), a hop length (stride) of 64 samples (16 ms), and a Hann window function.

The magnitude spectrogram is then mapped onto the Mel scale via a triangular filter bank with $n_{\mathrm{mel}} = 32$, distributing center frequencies according to the human auditory perception range. We present here the results based on Mel spectrograms because it yields better results than linear spectrograms. To ensure that extracted features focus on the target fish species while minimizing interference from high frequency and low frequency noise, each spectrogram is constrained within a predefined frequency range $80\text{Hz} \leq f \leq 1000\text{Hz}$. Mean normalization is applied to the feature matrix.

Since multiple fish vocalizations may be present simultaneously within a given time frame, each audio sample is labelled using a one-hot encoding scheme. We adopt a multi-label scheme to account for the possibility that multiple fish species may vocalize simultaneously

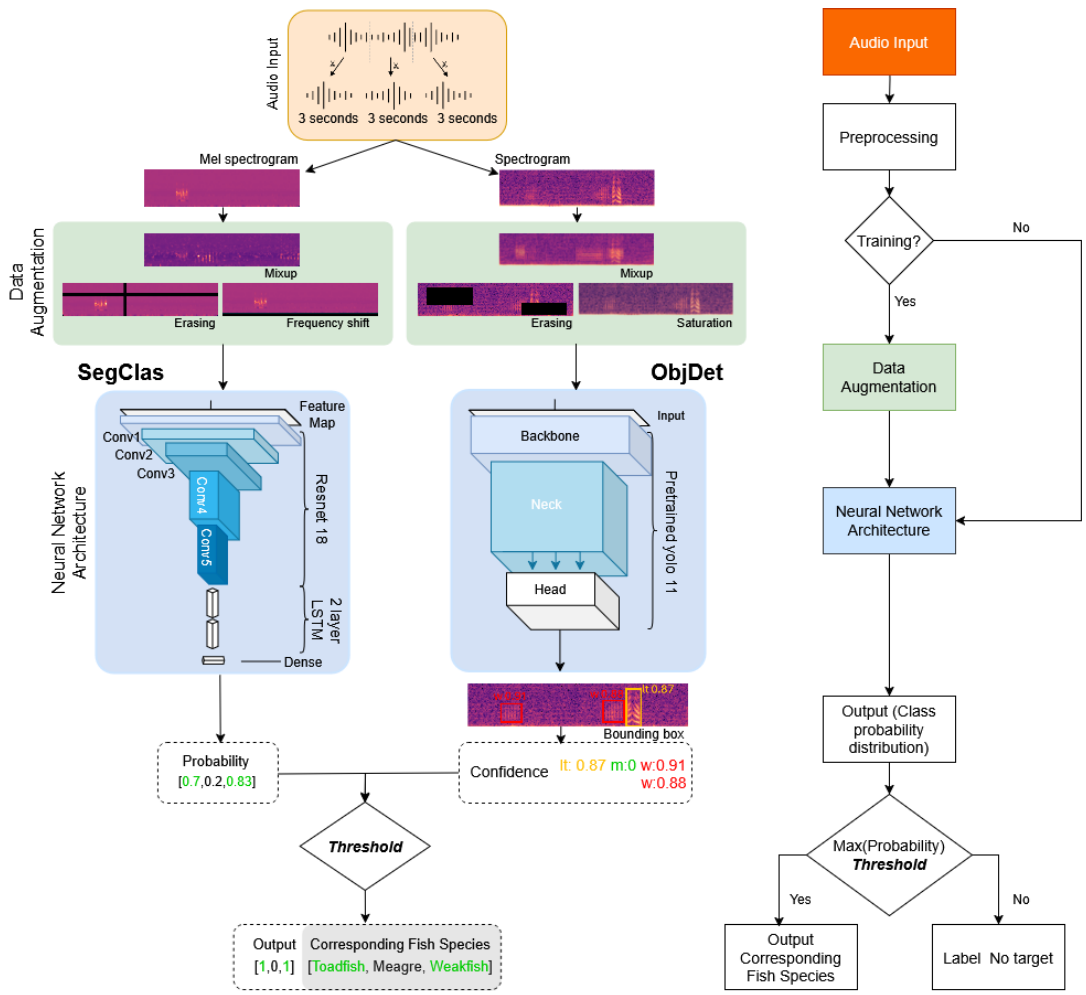


FIG. 2. Overview of the Integrated System Architecture. **Left:** Detailed schematic of the two processing pipelines. First, 4 kHz audio is divided into 3-second segments. In the SegClas pipeline, the audio is then normalized and converted into a Mel spectrogram before being fed into the neural network. Meanwhile, the ObjDet pipeline is fed an FFT-generated, linear-scaled spectrogram, which is represented as an RGB image. During training, data augmentation is randomly employed to enhance robustness. The final fish species classification is determined by thresholding the network's probability outputs. In the YOLO-based detection mechanism, object bounding boxes are produced with class label and confidence scores. A segment or box is classified as containing specific fish species sounds if the highest confidence exceeds a predefined threshold. **Right:** The complete system flowchart, illustrating the overall process from data preprocessing through to model training and to label generation.

within the same audio segment. For an audio sample x, we define its ground-truth label as a binary vector $Y = [y_1, y_2, \dots, y_C] \in \{0,1\}^C$ where C is the total number of fish species of interest, and $y_i = 1$ if species i is present in the segment, or 0 otherwise. This same labeling scheme is later applied to our YOLO-based approach(ObjDet) for consistency.

Data Augmentation After extracting features, we independently (or randomly) apply a range of data augmentation methods, ensuring also a portion of unmodified examples.

We use *time and frequency erasing* augmentation. These methods randomly zero out portions of the spectrogram along the temporal (Δt) and frequency (Δf) axes, simulating partial signal loss. Specifically, for each spectrogram, a frequency range $[f_1, f_2]$ and a time interval $[t_1, t_2]$ are randomly selected, where the erase widths are defined as $f_2 - f_1 = \Delta f$ and $t_2 - t_1 = \Delta t$, respectively. Erased regions are zeroed, as commonly done in audio and image augmentation, to simulate silence or partial dropout in a controlled manner, simulating missing information.

Frequency shift augmentation is applied to the Mel spectrogram features to improve the system's robustness against variations in the distance between a vocalizing fish, distortions in the acoustic sensor, and natural variations in fish calls. This choice is motivated by our observation that vocalizations from the same species may vary in frequency depending on biological and environmental factors. The shift magnitude is determined by a maximum and a minimum, frequency deviation f_{shift}, which is converted into a shift in Mel bins as follows: $S_{\max} = \frac{f_{\text{shift}}}{\Delta f}$, where Δf is the average frequency interval between adjacent Mel bins. Once the shift magnitude is determined, each Mel spectrogram is randomly shifted along the frequency axis using a circular shift operation: $\tilde{F} = \text{roll}(F, s, \dim = 2)$, where s is the randomly selected shift within the allowed range (max frequency shift is 100 Hz in this case), and the roll operation ensures that spectral content is smoothly adjusted without introducing artifacts.

Finally, we also apply *mixup* (Zhang, 2017) to further improve generalization and robustness to overlapping calls. In mixup, two training spectrograms (X_1, y_1) and (X_2, y_2) are linearly combined with a mixing ratio λ, drawn from a Beta distribution, to form: $X' = \lambda X_1 + (1-\lambda) X_2, \quad y' = \lambda y_1 + (1-\lambda) y_2$. Here, X' is the new spectrogram, and y' is the corresponding multi-label vector. By generating such interpolated examples, mixup simulates scenarios in which multiple fish vocalizations are partially blended, thus enhancing the model's robustness to noise and call overlap.

Network Architecture We adopt the lightweight 18-layer variant of the ResNet architecture (ResNet18) (He *et al.*, 2016) as the feature extractor, leveraging residual connections to help maintain stable gradients. Given an input Mel spectrogram X, the ResNet18 model extracts high-level feature representations, denoting F as the feature map $F = \text{ResNet18}(X)$.

While CNNs excel at capturing local spectral patterns, they are limited in modelling long-term temporal dependencies. To address this, we integrate LSTM (Yu *et al.*, 2019) after the ResNet18. The LSTM processes the sequence of feature embeddings F and learns the temporal relationships between different time frames: $H_t = \text{LSTM}(F_t, H_{t-1})$, where F_t is the feature representation at time step t, H_t is the hidden state of the LSTM at time step t, H_{t-1} is the hidden state from the previous time step.

Loss function For each audio sample, let $\hat{y} \in [0,1]^C$ denote the predicted probabilities. The Binary Cross-Entropy (BCE) Loss for each class is computed as: $\mathcal{L}_{\text{BCE}} = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$. To optimize the multi-label classification task, we adopt Focal Loss (Lin *et al.*, 2017) as loss function, which is particularly suited for handling class imbalance by down-weighting well-classified examples and focusing on hard-to-classify samples: $\mathcal{L}_{\text{Focal}} = \alpha(1-p_t)^\gamma \mathcal{L}_{\text{BCE}}$, where α is the balancing factor, γ is the focusing parameter and $p_t = \exp(-\mathcal{L}_{\text{BCE}})$ represents the probability of the correct class.

We additionally introduce frequency-based attention, label smoothing, and background sample weighting to focus on the relevant frequency bands, mitigate mislabelled examples, and appropriately handle recordings without fish vocalizations. Given a frequency activation matrix A extracted from the input spectrogram, the frequency weight vector is computed as: $W_{\text{freq}} = \frac{\max(0, AW_{\text{band}}^T)}{\sum_j \max(0, AW_{\text{band}}^T) + \epsilon}$, where W_{band} represents the predefined frequency band relevant for each class, and ϵ is a small constant to avoid division by zero. Specifically, we define W_{band} by highlighting the typical call frequency ranges of our target fish species (e.g., 50–600 Hz for toadfish), as identified in prior studies. This ensures that the loss function emphasizes the acoustically relevant bands for each species, reducing interference from out-of-range frequencies. The focal loss is then adjusted using this weight: $\mathcal{L}_{\text{weighted}} = W_{\text{freq}} \mathcal{L}_{\text{Focal}}$. Furthermore, to enhance model robustness, we apply label smoothing and incorporate background sample loss weighting to handle samples without identifiable fish vocalizations. Since some audio samples contain no identifiable fish vocalizations, we introduce parameter background loss weighting. Given that, a sample is labelled as no target if the sum of its ground-truth labels satisfies $\sum_{i=1}^{C} y_i < \epsilon$.

Finally, we define two separate loss terms: $\mathcal{L}_{\text{no\_target}} = W_{\text{bg}} \cdot \mathcal{L}_{\text{weighted}}, \quad \mathcal{L}_{\text{target}} = \mathcal{L}_{\text{weighted}}$. where W_{bg} is a predefined weight to emphasize background samples. Overall, the total loss is then computed as: $\mathcal{L}_{\text{total}} = a\mathcal{L}_{\text{target}} + b\mathcal{L}_{\text{no\_target}}$, where a and b are weighting coefficients to control the relative contribution of target and background losses.

Network Output The network ultimately outputs a probability p_c for each species c. How we convert these probabilities into binary labels (0 or 1) is described in Section II C.

2. Object Detection

Preprocessing Since ObjDet uses YOLO object detection, a different preprocessing scheme is applied. The 4 kHz audio is also cut into 3s segments but then it is converted into FFT-generated dB linear spectrogram with parameters *window size* = 256, *hop length* = 64, and *hann* window. These parameters optimize the differences between the visual patterns of calls produced by the species of interest. For ObjDet, Mel spectrograms produced inferior results; therefore, only results from linear spectrograms are presented. Frequencies above 1000 Hz are removed. Finally, as the underlying YOLO model is pre-trained on three-channel RGB images. We use Matplotlib's *Magma* colormap, as it most closely resembles Raven Pro's preferred colormap used for aural and visual inspection of the training and test datasets.

Data Augmentation In addition to time and frequency augmentation, for ObjDet, we employ erasing, mixup, and saturation shift.

Similarly to SegClas, ObjDet also employs erasing as a means of simulating missing information. This method randomly selects a rectangular patch of the image, defined by a frequency range $[f_1, f_2]$ and a time interval $[t_1, t_2]$, where the patch dimensions $\Delta f = f_2 - f_1$ and $\Delta t = t_2 - t_1$ are randomly chosen with the constraint that the erased region's area can only amount to 20% of the image. The erased region is set to zero, simulating missing information. The factor of 20% is not exceeded so it is highly unlikely that a fish sound would be completely erased.

ObjDet uses *mixup* (Zhang, 2017) to improve generalization, similar to the approach described for SegClas mixup (see section II B 1). Here, however, the input is a three-channel *dB spectrogram* (rather than a Mel spectrogram), mapped to RGB. Mixup combines two such spectrogram images (X_1, X_2) with labels (y_1, y_2) by drawing a mixing ratio λ from a Beta distribution, then forming an interpolated sample. During training, for each sample independently, there is a 20% chance that mixup is applied.

Lastly, *saturation shift* adjusts the brightness, contrast, and color properties of the images to simulate diverse signal to noise conditions. This can be defined by the parameter Δs (saturation shift) defined in a range of $[0, 20\%]$. For each augmentation, a random value is chosen from this range.

These augmentations are *not* applied to the training data before training begins, but just-in-time randomly for each training batch using RandAugment (Cubuk *et al.*, 2020).

Learning System We use the pre-trained YOLO version 11 nano (Khanam and Hussain, 2024) for our object detection-based approach. We choose this smallest variant in terms of parameters, as our dataset is relatively limited and contains only 3,848 training samples across three classes. The generated 3s segment spectrograms have a resolution of 188×64 pixels. Since YOLO requires input dimensions to be multiples of 32, we resize the spectrograms to 192×64 pixels.

C. Postprocessing

Fish calls can overlap, so we treat each 3-second segment as a multi-label classification problem: multiple species may appear simultaneously within the same audio clip. Both SegClas and ObjDet eventually produce a confidence score $\mathrm{conf}_c(x)$ for each species $c \in \{1, \ldots, C\}$.

In SegClas, this score ($\mathrm{conf}_c(x)$) is directly the network's probability output p_c. In ObjDet, the model is fine-tuned on our dataset using standard object detection, so its immediate outputs consist of bounding boxes accompanied by class labels and confidence scores. Note that while ObjDet is trained via standard object detection (bounding boxes + class scores), at inference time we aggregate per-class box confidences to generate a single segment-level score. To derive a multi-label classification for each 3-second spectrogram, we define: $\mathrm{conf}_c(x) = \max\{\text{box\_conf} \mid \text{box.class} = c\}$, where box_conf is the confidence associated with a bounding box labeled as species c. If no boxes are labeled with species c, then $\mathrm{conf}_c(x) = 0$.

To convert these continuous scores into binary presence/absence predictions $\hat{y}_c \in \{0, 1\}$, we apply a thresholding step: $\hat{y}_c = 1 \text{if } \mathrm{conf}_c(x) \geq T_c$, and 0 otherwise. Because this is a multi-label setting, multiple $\hat{y}_c$ may be 1 (i.e., multiple fish species can co-occur).

Instead of applying a single universal threshold across all classes, we determine a separate T_c for each species by maximizing the F1 score (see Section II D) on a validation split: $T_c = \operatorname{argmax} F1(\hat{Y}_{c,\tau}, Y_c), \tau \in [0, 1]$ where $F1$ measures precision–recall balance between ground-truth Y_c and the model predictions $\hat{Y}_{c,\tau}$, binarized at threshold τ. This class-specific approach accommodates differences in call duration, signal-to-noise ratio, and background interference among the target fish species.

D. Evaluation Metrics

The evaluation metrics used in this study are standard for multi-label classification and object detection tasks. These metrics were computed on the test dataset described in Section II A. Metrics were assessed using a hybrid cross-validation approach to ensure an independent test set evaluation, with the test set remaining unseen during training. The training data underwent 5-fold cross-validation, producing five models, each evaluated on the fixed test set to provide a robust performance estimate.

Classification Metrics We evaluated the performance using precision, recall, F1 score, and accuracy as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN},$$
$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP, FN are true positive, true negative, false positive, and false negative, respectively.

A high *precision* indicates that most segments classified as fish indeed contain fish calls (i.e., good correctness), but does not reveal how many fish calls were missed. In contrast, a high *recall* signifies that the system has found a large proportion of the actual fish calls (i.e., good coverage), but does not indicate how many non-fish segments were incorrectly labeled. Consequently, precision and recall must both be considered to fully understand a detector's performance, which is why the *F1 score* combines them into a single measure (ranging from 0 to 1). Finally, *accuracy* measures the fraction of segments (fish or non-fish) correctly labeled overall.

Additionally, we computed subset accuracy (also known as exact match), which measures the proportion of segments where the predicted and ground-truth label sets match exactly: Subset Acc. $= \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)$, where $\hat{y}_i$ is the predicted label set and y_i is the ground-truth label set for sample i, and $\mathbb{I}(\cdot)$ is the indicator function which is 1 when $\hat{y}_i = y_i$, and 0 otherwise. Note that this stricter metric penalizes any incorrect species label in a segment, making it harder to achieve high subset accuracy when multiple fish may vocalize simultaneously.

For each fish species, we computed precision, recall, F1 score, and accuracy in a one-vs-all manner, that is, assuming a binary classification for each species. Additionally, F1 score, accuracy, and subset accuracy were calculated in full. All metrics were calculated for both SegClas and ObjDet approaches.

Count Estimation Metrics To assess count estimation accuracy, we calculated Relative Error (RE) for each class: $\mathrm{RE}_c = \frac{|\hat{n}_c - n_c|}{n_c}$, where $\hat{n}_c$ and n_c are the predicted and ground-truth segment-based counts for class c, respectively. Note that we can use two counting approaches: Segment-based count (default method)—Each classified segment contributes one count toward the corresponding species. This approach provides a more stable estimate, particularly in dense choruses where individual vocalizations may be indistinguishable. Total count (ObjDet only)—Counts every detected vocalization based on the number of bounding boxes assigned to a species.

Inference Time quantifies the average processing time for a 3-second audio segment, including spectrogram generation and classifier inference. This is reported as an average over the entire test dataset to assess computational efficiency. Measurements were taken on a 2023 M2 MacBook Pro with 16 GB of memory running MacOS version 15.3.2 and using Apple's Metal Performance Shaders (MPS) framework in PyTorch 2.5.1 to speed up inference.

III. RESULTS

In this section we evaluate the performance of SegClas and ObjDet models for marine species vocalization detection across multiple fish species, comparing their accuracy metrics, classification capabilities, and temporal prediction patterns through statistical analysis and visual examples. The thresholds are chosen as optimizers of the F1 score on the training data. For SegClas the optimal thresholds according to the training data are $(0.6, 0.56, 0.54)$ for (lt, m, w), respectively, while the optimal thresholds for ObjDet are $(0.25, 0.66, 0.25)$. These thresholds apply to the test data in the following results. See supplemental materials for details on this choice. A demonstration, data, and code is available at https://github.com/NabiaAI/Argyrosomus.

The evaluation results on the test data are summarized in Table I by the mean and standard deviation of the five folds. ObjDet generally outperforms SegClas by a small margin, with an F1 score of 92.0%, accuracy of 97.4%, and subset accuracy of 92.8%, compared to 87.6%, 96.5%, and 90.3%, respectively. For meagre classification, both models excel in different metrics, with SegClas producing much lower Count RE compared to ObjDet. For the weakfish, we also observe very high scores across most metrics in both approaches ($\geq 85\%$), except for the F1 score and recall in SegClas. In the weakfish we observe the highest count RE. For the toadfish, Count RE, ObjDet excels by a small margin on all metrics. Lastly, the inference time was also measured, SegClas is roughly 25% faster than ObjDet.

Figure 3 compares the multi-label confusion matrices of both models. This details where the models are especially prone to misclassifications. Both models seem to particularly misclassify the segments with only toadfish as noise (102 (10.6%) incorrect segments for SegClas, 57 (5.9%) incorrect for ObjDet), the segments with both toadfish and meagre simultaneously as only meagre (62 (28.8%) SegClas, 38 (17.7%) ObjDet), and the segments with toadfish and weakfish as only weakfish (41 (41.8%) SegClas, 25 (25.5%) ObjDet). For the remaining pairings, the numbers for SegClas and ObjDet are also very similar (difference < 15). Notably, the test dataset included a high proportion of non-fish segments, reflecting realistic conditions in continuous recordings. Both models maintained a true negative rate above 95%, indicating reliable discrimination between fish and non-fish sounds.

TABLE I. Comparison of SegClas and ObjDet on test dataset. Given numbers are means with their standard deviations from repeated testing on a fixed test set using 5 fold cross-validated training sets. The best results are <u>underlined</u>.

Metric	**SegClas**	**ObjDet**
F1 Score [%]	87.4 ± 1.3	<u>92.0 ± 1.5</u>
Accuracy [%]	96.6 ± 0.9	<u>97.4 ± 1.0</u>
Subset Acc. [%]	90.1 ± 1.1	<u>92.8 ± 1.2</u>
Inference Time per Segment [ms]	<u>1.22</u>	1.63
Meagre		
F1 Score [%]	93.3 ± 1.0	<u>94.2 ± 1.2</u>
Precision [%]	95.1 ± 0.9	<u>98.8 ± 0.8</u>
Recall [%]	<u>91.7 ± 1.1</u>	90.0 ± 1.4
Accuracy [%]	<u>98.1 ± 0.8</u>	98.0 ± 0.9
Count RE [%]	<u>4.3 ± 0.7</u>	9.8 ± 1.1
Weakfish		
F1 Score [%]	82.7 ± 1.6	<u>90.2 ± 1.8</u>
Precision [%]	92.1 ± 1.1	<u>96.4 ± 1.0</u>
Recall [%]	75.2 ± 1.9	<u>84.7 ± 2.2</u>
Accuracy [%]	<u>98.2 ± 1.2</u>	98.0 ± 0.8
Count RE [%]	22.3 ± 1.5	<u>13.9 ± 1.9</u>
L. Toadfish		
F1 Score [%]	86.6 ± 1.5	<u>91.5 ± 1.7</u>
Precision [%]	92.8 ± 1.2	<u>95.4 ± 1.3</u>
Recall [%]	81.8 ± 1.8	<u>88.0 ± 2.0</u>
Accuracy [%]	93.5 ± 1.0	<u>95.1 ± 1.1</u>
Count RE [%]	13.7 ± 1.1	<u>8.4 ± 0.9</u>

Figure 4 shows qualitative examples of the inference, object detection, and classification process. Both models effectively identify a wide range of vocalizations across various signal-to-noise ratios (Figure 4a)), but can fail when dealing with extremely low signal-to-noise ratio vocalizations (Figure 3b)). Examples like Figure 4b) illustrate instances where sections containing both toadfish and weakfish sounds are labeled solely as weakfish. Meagre choruses (Figure 4c)) as well as overlapping sounds (Figure 4a)) are well recognized. Other mistakes include background noise that resembles low signal-to-noise ratio meagre calls (Figure 4d)), and low-frequency noise which is misclassified as toadfish (Figure 4e)). Furthermore, distinguishing weakfish grunts within meagre knocking sounds is generally challenging for both models (Fig-

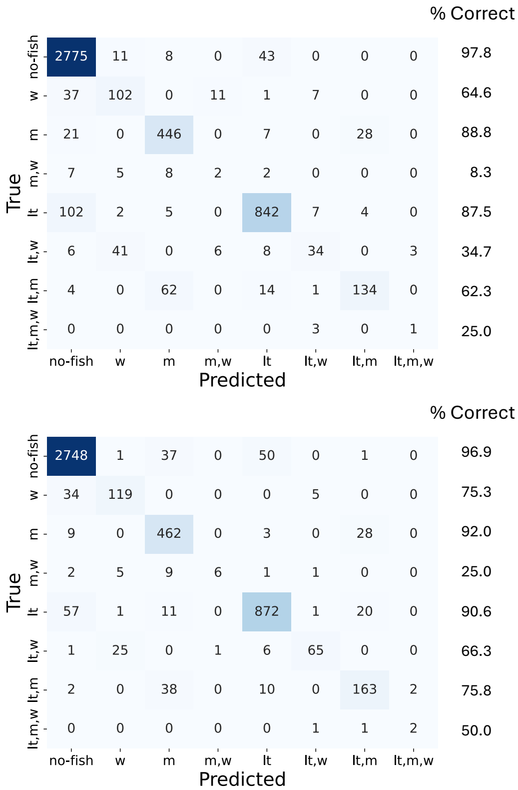
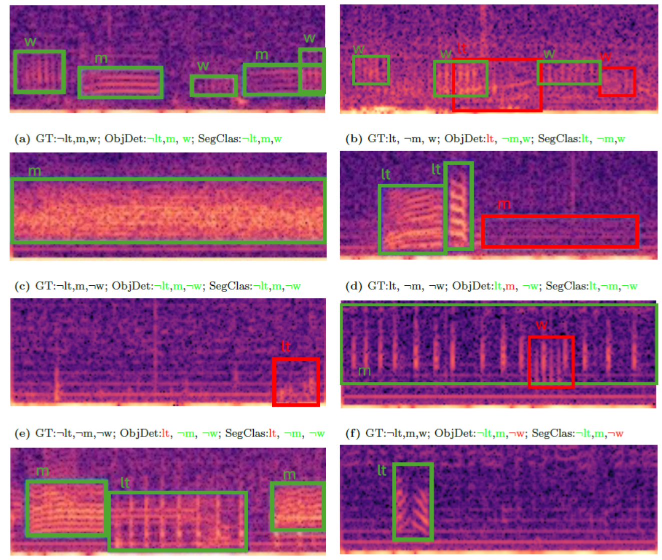


FIG. 3. Multi-label confusion matrices of SegClas (top) and ObjDet (bottom) on test dataset. The numbers represent total 3s-segment classifications of the best model of each approach. lt - Lusitanian toadfish, m - meagre, w - weakfish.



(a) GT:¬lt,m,w; ObjDet:¬lt,m, w; SegClas:¬lt,m,w
(b) GT:lt, ¬m, w; ObjDet:lt, ¬m,w; SegClas:lt, ¬m,w
(c) GT:¬lt,m,¬w; ObjDet:¬lt,m,¬w; SegClas:¬lt,m,¬w
(d) GT:lt, ¬m, ¬w; ObjDet:lt,m, ¬w; SegClas:lt,¬m,¬w
(e) GT:¬lt,¬m,¬w; ObjDet:lt, ¬m, ¬w; SegClas:lt, ¬m, ¬w
(f) GT:¬lt,m,w; ObjDet:¬lt,m,¬w; SegClas:¬lt,m,¬w
(g) GT:lt,m,¬w; ObjDet:lt,m, ¬w; SegClas:¬lt,m,¬w
(h) GT:lt,¬m,¬w; ObjDet:lt,¬m,¬w; SegClas:¬lt,m,¬w

FIG. 4. Qualitative examples of the inference, object detection, and classification process. The shown spectrograms of 3 s audio segments are annotated with bounding boxes by ObjDet. Underneath the annotated spectrograms, the classifications are given by the ground-truth (GT), by ObjDet, and by SegClas. Green indicates correct boxes or classifications, while red indicates incorrect ones. Spectrograms were produced with fast Fourier transform (FFT) = 256, frequency in a linear scale from 0 up to 1 kHz, and a time frame of 3 s.

ure 4f)). In the test dataset, many errors in segments containing both meagre and weakfish calls occur within these knocking sounds, where such sections are often misidentified as meagre only or as no-fish (see Figure 3). Furthermore, distinguishing rarer toadfish sounds, such as the double-croak, was occasionally challenging for SegClass (Figure 4g), Figure 4h)).

Figure 5-a) shows a comparison of both models on the test data in more detail. Additionally, for ObjDet, the number of total vocalizations (as opposed to segment-based vocalizations) is shown in Figure 5-b). Both figures give the date and point in time of the 10 min interval the counts were accumulated in. During winter, when fish vocalizations are nearly absent, the number of detections is correspondingly low. Both models show the largest discrepancy on April 18, 2021 at 4 pm for weakfish. On the same interval, meagre performance is also off. Visual inspection revealed that this time interval features high noise levels (electro-static noise and a boat passing by). On July 6, 2021 the largest discrepancy for Toadfish is observed. When comparing Figure 5-a) to Figure 5-b), it is evident that Figure 5-b) exhibits smaller relative errors. However, additional discrepancies were found in Figure 5-b) on July 24, 2017, with approximately 100 false negatives for toadfish calls.

Additionally, Figure 6 shows predicted segment-based counts of both models on 24-hour continuous recordings from April 19, 2017 and April 27, 2021, alongside corresponding long-term spectrograms. Figure 6-a) shows a distinct meagre chorus between 4 pm and 10 pm. The predictions of both models are essentially the same during this phase. ObjDet predicts virtually no weakfish sound occurrence, which aligns with the known weakfish activity patterns for this day. However, SegClas incorrectly predicts two smaller peaks of weakfish activity at 4 pm and 9 pm, raising concerns about the model's false positives. On the other hand, both models predict toadfish activity throughout the day except for a break during the peak of the meagre chorus around 6 pm. ObjDet generally reports higher levels of toadfish activity, which is consistent with the higher performance of this approach (see Table 1). In 2021 (Figure 6-b), both models are well-aligned in detecting the chorus activity of toadfish, meagre and weakfish. However, SegClas predicts slightly higher activity of weakfish from 12 am to 2 pm which is incorrect.

IV. DISCUSSION

Our study demonstrates the potential of using deep learning to analyze Passive Acoustic Monitoring (PAM) data, highlighting the feasibility of large-scale, non-invasive assessment of fish populations in complex estuarine environments. The high classification accuracy and F1 scores of both deep learning frameworks

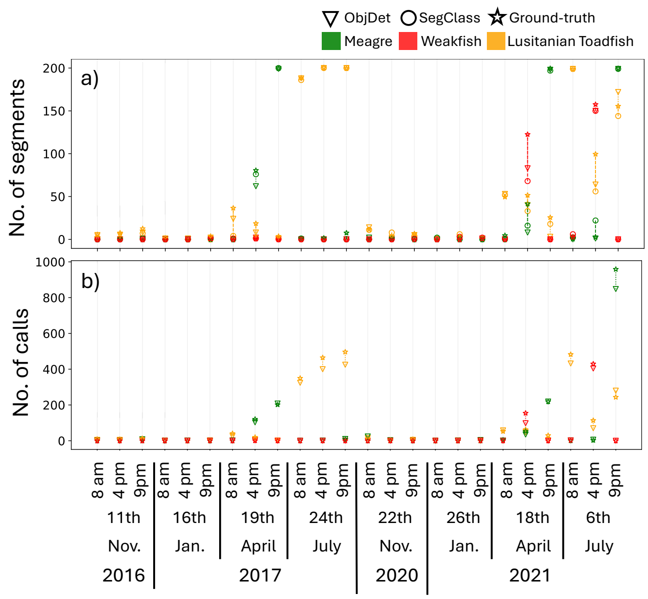


FIG. 5. Comparison of performance on test data between ObjDet, SegClas, and ground-truth (GT). Figure 5-a) is showing number of 3 s segments containing fish vocalization over the time and date of the 10 min interval from which the count originates. Figure 5-b) (only for ObjDet) is showing total number of fish vocalizations over the same 10 min intervals as Figure 5-a). Dotted lines indicate the difference of ObjDet to gt and dashed lines indicate the difference of SegClas to gt.

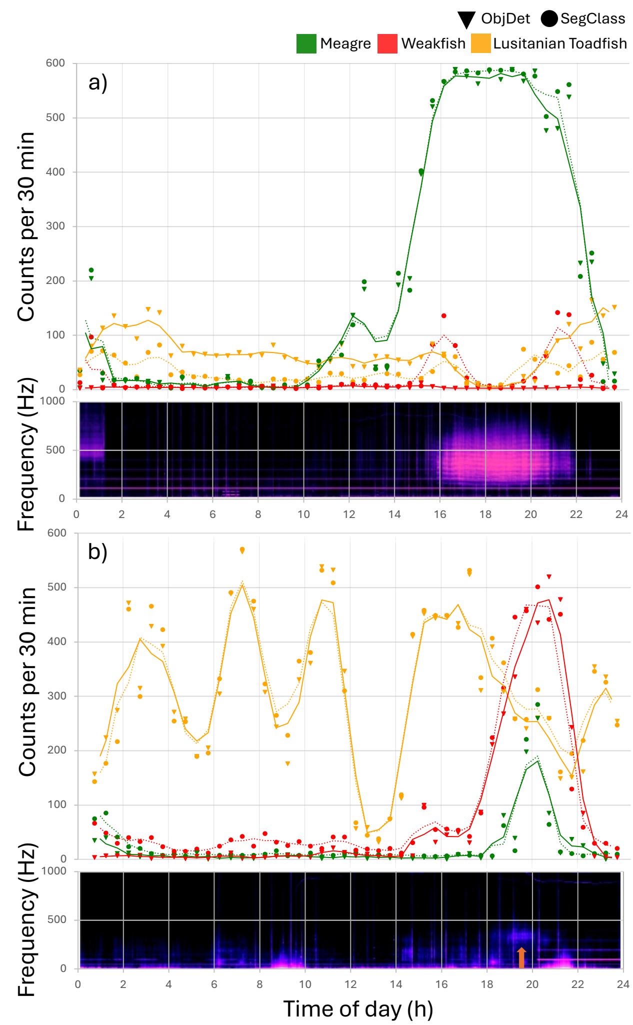


FIG. 6. Predicted fish calling activity using ObjDet (solid lines, triangle markers) and SegClas (dotted lines, circle markers) over 24 hours. For this figure, the segment-based count interval is 30 min. Long-term spectrogram (FFT sampling rate 1024, hop length 512, window type Hann; averaged over 1 min bins) are shown for April 19, 2017 (a) and April 27, 2021 (b). The lines represent moving averages, and the orange arrow (b spectrogram) marks the peak of calling activity for the meagre and weakfish on that day.

validate their effectiveness in identifying and quantifying fish vocalizations, even under challenging acoustic conditions. Moreover, comparing object detection and classification-based approaches provides valuable insights into their respective strengths and trade-offs, informing future methodological choices based on ecological and operational needs. These findings emphasise the growing role of artificial intelligence in ecoacoustic monitoring, contributing to improve biodiversity conservation and resource management strategies.

A. Model Performance and Technical Insights

We implemented and compared two deep learing approaches for detecting and classify multiple sounds of multiple fish species: SegClas and ObjDet. SegClas is a classification-based method that assigns labels to entire audio segments, whereas ObjDet is an object detection-based approach that identifies specific vocal events within spectrograms. ObjDet can provide more accurate annotation and higher accuracy and F1 scores. This advantage stems from its ability to detect individual vocal events using YOLO's bounding-box mechanism, which enhances precision (Redmon *et al.*, 2016). On the other hand, SegClas offers a significant reduction in inference time—approximately 25% faster than ObjDet—and requires only segment-level labels rather than detailed bounding-box annotations. This aligns with previous findings that reducing annotation overhead is crucial for large-scale ecological surveys (Demir *et al.*, 2020; Jung *et al.*, 2021). Even with high-performing detection models, the burden of detailed labeling can be a limiting factor. By contrast, SegClas optimizes both speed and practicality, making it a compelling choice for real-time monitoring applications where computational resources are constrained.

Our findings align with recent studies in marine bioacoustics that have also applied ResNet CNNs for fish sound detection and classification. For instance, Waddell *et al.* (2021) used a pre-trained ResNet-50 for call-type classification of fish sounds, achieving F1 scores ranging from 0.44 to 0.77 across six call types. Compared to our results, these lower scores may be attributed to the challenges of multi-class classification, mainly when limited manual annotations are available for certain sound types. Similarly, (Munger *et al.*, 2022) achieved an F1 score of 0.86 for classifying damselfish sounds using a ResNet-50 CNN, demonstrating strong performance, albeit slightly lower than ours. Additionally, Mouy *et al.* (2024) employed a ResNet CNN to distinguish between "fish" and "non-fish" sounds in a dataset of unidentified fish sounds, achieving an F1 score of 0.82. Notably, their study used shorter 0.2-second segments, which was appropriate given the predominant fish sounds in their dataset. In contrast, (Waddell *et al.*, 2021) used 0.5-second segments, (Munger *et al.*, 2022) used 2-second segments, and our study employed 3-second segments.

Other machine learning techniques have also been applied to annotated fish sounds. For instance, Malfante *et al.* (2018) employed both random forest and support vector machines to classify six fish call types, achieving F1 scores exceeding 0.90 and accuracies up to 96.9%. Noda *et al.* (2016) achieved an impressive F1 score of 0.98 using an SVM-based algorithm for classifying sounds from 128 fish species. However, their work was limited by a smaller dataset that did not include noise recordings, which could affect generalizability in real-world applications. In fact, preliminary test in our study was restricted to a small dataset with less variability and without noise, and showed higher performance. Mouy *et al.* (2024) employed a combination of detection of acoustic transients in the spectrogram and the classification using Random Forest, achieving a low F1 score of 0.43 (against the F1 score of 0.82 obtained with ResNET on the same datasets).

More relevantly, several studies have applied sound classification techniques to species examined in our study, as well as other members of the Sciaenidae family, which includes the meagre and the weakfish. For instance, Vieira *et al.* (2015) developed the first fish sound classification system, which was based on hidden Markov models (HMMs) and applied to Lusitanian toadfish vocalizations. This study focused on identifying sounds produced by individual males and also introduced a call-type recognition system. While it achieved high recall ($> 90\%$) for the boatwhistle call type, it struggled with other call types produced by the species. In contrast, our study did not differentiate between specific call types. However, both SegClas and ObjDet successfully classified most call types as belonging to the Lusitanian toadfish. This suggests that future applications could leverage CNNs to detect and distinguish between the various calls of this species, which has a well-documented, diverse vocal repertoire Amorim *et al.* (2008). A key advantage of Vieira *et al.* (2015) technique is its ability to annotate individual calls, enabling the extraction of ecologically relevant features (e.g., call duration). However, it struggled with overlapping calls. In our study, the ObjDet approach, using YOLO's bounding-box mechanism, offers the same advantage while demonstrating greater accuracy in identifying overlapping calls. Vieira *et al.* (2019) also employed an HMM-based system to detect and classify meagre calls over seven months of continuous data recorded in captivity. While the system effectively tracked calls of interest with 78% accuracy, it faced challenges in classifying sounds based on predefined categories. This difficulty sparked a discussion on the proper definition of the true call types of this species (Bolgan *et al.*, 2020). Using this same technique, a system was also successfully employed to track the choruses produced by meagre in the wild over a four-year period, achieving an accuracy of 96.7%. However, it was not designed to distinguish between the sounds of this species and those of the newly invasive weakfish (Vieira *et al.*, 2022). Overall, both SegClas and ObjDet performed well in handling choruses with these two species, however SegClas exhibited false positives for weakfish when only meagre choruses were present (see Figure 6). While scienid species are known to produce continuous choruses on certain days (Vieira *et al.*, 2022), our recordings were predominantly dominated by meagre, likely because weakfish schools are usually positioned farther from the hydrophone (Matos *et al.*, 2024). In other locations, where both species might produce overlapping continuous choruses, their differentiation may pose different challenges. Other systems, such as support vector machines (SVM), k-nearest neighbors (k-NN), periodicity-coded non-negative matrix factorization (PC-NMF), and Gaussian mixture models (GMM), as well as simpler sound detectors, have also been applied to the analysis of sounds and choruses in other sciaenid species, as well as in choruses from other families (Harakawa *et al.*, 2018; Hawkins *et al.*, 2025; Lin *et al.*, 2018; Monczak *et al.*, 2019; Siddagangaiah *et al.*, 2019).

B. Biological and Ecological Implications

From an ecological viewpoint, both methods demonstrate potential for long-term PAM in dynamic estuarine systems such as the Tagus estuary. This location features overlapping calls from Lusitanian toadfish, meagre, and weakfish, often intertwined with environmental and anthropogenic noises (Amorim *et al.*, 2023; Vieira *et al.*, 2021a). Effective discrimination among these taxa is fundamental for understanding their spatiotemporal distribution, mating periods and spawning sites, and general habitat usage (Lindseth and Lobel, 2018). For instance, accurate detection of chorusing behaviors can inform peak spawning windows, aiding marine managers in determining critical conservation periods or adjusting fishery regulations (McWilliam *et al.*, 2017; Stratoudakis *et al.*, 2024). Additionally, monitoring invasive species like the weakfish can prompt adaptive management interventions (Amorim *et al.*, 2023; Lodge *et al.*, 2016; Stratoudakis *et al.*, 2024).

The SegClas approach, with its lighter annotation requirements, facilitates swift deployment in regions where extensive bounding-box labeling is unfeasible. Ecologically, this allows researchers to expand monitoring to multiple sites and gather broad-scale temporal data on fish assemblages. Meanwhile, ObjDet addresses scenarios where precise call localization is essential—for example, studying fine-scale interactions between co-occurring species, quantifying vocalization rates within choruses, or investigating how anthropogenic disturbances (e.g., boat traffic) may affect fish call dynamics that can underlie reproductive success. (Vieira *et al.*, 2022, 2024). Although ObjDet offers higher accuracy in many metrics, the added computational cost and annotation effort may limit its adoption in resource-constrained projects. Thus, the best method depends on balancing logistical constraints (e.g., labeling budget, hardware capacity) against ecological questions of interest.

C. Remaining Challenges and Future Directions

Despite robust data augmentation (time-frequency erasing, frequency shifting, mixup), certain classification errors persisted. Overlapping calls among acoustically similar species remain a bottleneck, particularly in dense choruses (Gibb *et al.*, 2019). Further refinements could involve *hybrid architectures*, merging the localization strengths of ObjDet with the simpler segmentation pipeline of SegClas. For example, a two-stage strategy might first label coarse segments to identify candidate fish presence and then apply a lighter object detector for precise bounding-box proposals within segments flagged as "active."

Another promising direction is *adaptive thresholding* or region-specific threshold tuning based on local acoustic conditions. For example, by incorporating environmental metadata (e.g., tide levels, salinity, known diurnal cycles) in deep learning frameworks, thresholds might be dynamically adjusted to accommodate shifting noise floors and species-specific call patterns. Additionaly, integrating *domain adaptation* techniques could further improve performance in unstudied or evolving underwater environments, such as those affected by climate-driven habitat shifts.

D. Conclusion

This study highlights the potential of advanced deep learning methods in tackling complex underwater soundscapes for the assessments of soniferous fish. By comparing a segmentation-based CNN–LSTM framework SegClas with a YOLO-based object detection model ObjDet, we reveal tangible trade-offs between accuracy, labeling cost, computational overhead, and interpretability. From a bioacoustic perspective, both methods have demonstrated efficacy in isolating fish vocalizations amid potentially challenging real conditions and overlapping calls, thus providing a non-intrusive approach to monitor critical life-history events and habitat usage.

SegClas proves advantageous for broad, long-term surveys where minimal annotation and rapid inference are paramount. ObjDet offers finer-grained call localization and improved recognition in complex conditions, albeit at the expense of extensive bounding-box labeling and higher inference times. Ultimately, choosing between these approaches depends on study objectives—whether the emphasis is on precise call-by-call analyses or on large-scale continuous monitoring.

Continued refinement of these systems, along with adaptive thresholding and hybrid modeling strategies, will further mitigate misclassifications and expand their suitability in diverse ecological contexts. As passive acoustic monitoring becomes more integrated into ecosystem management, these deep learning frameworks have the potential to provide real-time insights. By processing and transmitting data from field stations in real time, these systems could facilitate early detection of ecological threats, supporting the protection and sustainable management of marine environments.

V. ACKNOWLEDGMENTS

We thank the Portuguese Air Force Base No. 6 for allowing the collection of the acoustic data. This work was financed by the Fundação para a Ciência e a Tecnologia (FCT) through project UID/04292-Centro de Ciências do Mar e do Ambiente, awarded to MARE; project LA/P/0069/2020 (https://doi.org/10.54499/LA/P/0069/2020) granted to the Associate Laboratory ARNET; UID/00329/2025 awarded to CE3C; UIDB/50021/2020 to INESC-ID; and projects Relevant (PTDC/CCI-COM/5060/2021) and ITI/LARSyS (LA/P/0083/2020, UIDP/50009/2020). Additional support was provided by the CoastNet Infrastructure and the respective projects (MAR-016.9.1-FEAMPA-00010 and LISBOA2030-FEDER-01319200), as well as by the Horizon Europe BIG (GA 952226), and Bauhaus of the Seas (GA 101079995).

VI. SUPPLEMENTARY MATERIAL

See supplementary material at [URL will be inserted by AIP].

[1] https://www.ravensoundsoftware.com/

Amorim, M. C. P., Simões, J. M., and Fonseca, P. J. (**2008**). "Acoustic communication in the lusitanian toadfish, halobatrachus didactylus: evidence for an unusual large vocal repertoire," Journal of the Marine Biological Association of the United Kingdom **88**(5), 1069–1073.

Amorim, M. C. P., Wanjala, J. A., Vieira, M., Bolgan, M., Connaughton, M. A., Pereira, B. P., Fonseca, P. J., and Ribeiro, F. (**2023**). "Detection of invasive fish species with passive acoustics: discriminating between native and non-indigenous sciaenids," Marine Environmental Research **188**, 106017.

Boelman, N. T., Asner, G. P., Hart, P. J., and Martin, R. E. (**2007**). "Multi-trophic invasion resistance in hawaii: bioacoustics, field surveys, and airborne remote sensing," Ecological Applications **17**(8), 2137–2144.

Bolgan, M., Parmentier, E., Picciulin, M., Hadjioannou, L., and Di Iorio, L. (**2023**). "Use of passive acoustic monitoring to fill knowledge gaps of fish global conservation status," Aquatic Conservation: Marine and Freshwater Ecosystems **33**(12), 1580–1589.

Bolgan, M., Pereira, B. P., Crucianelli, A., Mylonas, C. C., Pousão-Ferreira, P., Parmentier, E., Fonseca, P. J., and Amorim, M. C. P. (**2020**). "Vocal repertoire and consistency of call features in the meagre argyrosomous regius (asso, 1801),” Plos one **15**(11), e0241792.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (**2020**). “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703.

Demir, F., Abdullah, D. A., and Sengur, A. (**2020**). “A new deep cnn model for environmental sound classification,” IEEE Access **8**, 66529–66537.

Gibb, R., Browning, E., Glover-Kapfer, P., and Jones, K. E. (**2019**). “Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring,” Methods in Ecology and Evolution **10**(2), 169–185.

Guyot, P., Alix, F., Guerin, T., Lambeaux, E., and Rotureau, A. (**2021**). “Fish migration monitoring from audio detection with cnns,” in *Proceedings of the 16th International Audio Mostly Conference*, pp. 244–247.

Harakawa, R., Ogawa, T., Haseyama, M., and Akamatsu, T. (**2018**). “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 **144**(5), 2709–2718.

Hawkins, L. A., Parsons, M. J., McCauley, R. D., Parnum, I. M., and Erbe, C. (**2025**). “Passive acoustic monitoring of fish choruses: a review to inform the development of a monitoring and management tool,” Reviews in Fish Biology and Fisheries 1–28.

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Jiang, P., Ergu, D., Liu, F., Cai, Y., and Ma, B. (**2022**). “A review of yolo algorithm developments,” Procedia computer science **199**, 1066–1073.

Jung, D.-H., Kim, N. Y., Moon, S. H., Jhin, C., Kim, H.-J., Yang, J.-S., Kim, H. S., Lee, T. S., Lee, J. Y., and Park, S. H. (**2021**). “Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering,” Animals **11**(2), 357.

Khanam, R., and Hussain, M. (**2024**). “Yolov11: An overview of the key architectural enhancements,” arXiv preprint arXiv:2410.17725 .

Kvsn, R. R., Montgomery, J., Garg, S., and Charleston, M. (**2020**). “Bioacoustics data analysis–a taxonomy, survey and open challenges,” IEEE Access **8**, 57684–57708.

Lai, G., Chang, W.-C., Yang, Y., and Liu, H. (**2018**). “Modeling long-and short-term temporal patterns with deep neural networks,” in *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104.

Lin, T.-H., Tsao, Y., and Akamatsu, T. (**2018**). “Comparison of passive acoustic soniferous fish monitoring with supervised and unsupervised approaches,” The Journal of the Acoustical Society of America **143**(4), EL278–EL284.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (**2017**). “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Lindseth, A. V., and Lobel, P. S. (**2018**). “Underwater soundscape monitoring and fish bioacoustics: a review,” Fishes **3**(3), 36.

Lodge, D. M., Simonin, P. W., Burgiel, S. W., Keller, R. P., Bossenbroek, J. M., Jerde, C. L., Kramer, A. M., Rutherford, E. S., Barnes, M. A., Wittmann, M. E. *et al.* (**2016**). “Risk analysis and bioeconomics of invasive species to inform policy and management,” Annual Review of Environment and Resources **41**(1), 453–488.

Malfante, M., Mars, J. I., Dalla Mura, M., and Gervaise, C. (**2018**). “Automatic fish sounds classification,” The Journal of the Acoustical Society of America **143**(5), 2834–2846.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., and Tyack, P. L. (**2013**). “Estimating animal population density using passive acoustics,” Biological reviews **88**(2), 287–309.

Matos, A. B., Vieira, M., Amorim, M. C. P., and Fonseca, P. J. (**2024**). “Reaction of two sciaenid species to passing boats: Insights from passive acoustic localisation,” Estuarine, Coastal and Shelf Science **311**, 109012.

McWilliam, J. N., McCauley, R. D., Erbe, C., and Parsons, M. J. (**2017**). “Patterns of biophonic periodicity on coral reefs in the great barrier reef,” Scientific Reports **7**(1), 17459.

Monczak, A., Ji, Y., Soueidan, J., and Montie, E. W. (**2019**). “Automatic detection, classification, and quantification of sciaenid fish calls in an estuarine soundscape in the southeast united states,” PloS one **14**(1), e0209914.

Mouy, X., Archer, S. K., Dosso, S., Dudas, S., English, P., Foord, C., Halliday, W., Juanes, F., Lancaster, D., Van Parijs, S. *et al.* (**2024**). “Automatic detection of unidentified fish sounds: a comparison of traditional machine learning with deep learning,” Frontiers in Remote Sensing **5**, 1439995.

Munger, J. E., Herrera, D. P., Haver, S. M., Waterhouse, L., McKenna, M. F., Dziak, R. P., Gedamke, J., Heppell, S. A., and Haxel, J. H. (**2022**). “Machine learning analysis reveals relationship between pomacentrid calls and environmental cues,” Marine Ecology Progress Series **681**, 197–210.

Noda, J. J., Travieso, C. M., and Sánchez-Rodríguez, D. (**2016**). “Automatic taxonomic classification of fish based on their acoustic signals,” Applied Sciences **6**(12), 443.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (**2016**). “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Ribeiro, Jr, J. W., Harmon, K., Leite, G. A., de Melo, T. N., LeBien, J., and Campos-Cerqueira, M. (**2022**). “Passive acoustic monitoring as a tool to investigate the spatial distribution of invasive alien species,” Remote Sensing **14**(18), 4565.

Rippel, O., Snoek, J., and Adams, R. P. (**2015**). “Spectral representations for convolutional neural networks,” Advances in neural information processing systems **28**.

Siddagangaiah, S., Chen, C.-F., Hu, W.-C., and Pieretti, N. (**2019**). “A complexity-entropy based approach for the detection of fish choruses,” Entropy **21**(10), 977.

Stratoudakis, Y., Vieira, M., Marques, J. P., Amorim, M. C. P., Fonseca, P. J., and Quintella, B. R. (**2024**). “Long-term passive acoustic monitoring to support adaptive management in a sciaenid fishery (tagus estuary, portugal),” Reviews in Fish Biology and Fisheries **34**(1), 491–510.

Van Hoeck, R. V., Paxton, A. B., Bohnenstiehl, D. R., Taylor, J. C., Fodrie, F. J., and Peterson, C. H. (**2021**). “Passive acoustic monitoring complements traditional methods for assessing marine habitat enhancement outcomes,” Ecosphere **12**(11), e03840.

Vieira, M., Amorim, M. C. P., and Fonseca, P. J. (**2021**a). “Vocal rhythms in nesting lusitanian toadfish, halobatrachus didactylus,” Ecological Informatics **63**, 101281.

Vieira, M., Amorim, M. C. P., Marques, T. A., and Fonseca, P. J. (**2022**). “Temperature mediates chorusing behaviour associated with spawning in the sciaenid argyrosomus regius,” Marine Ecology Progress Series **697**, 109–124.

Vieira, M., Amorim, M. C. P., Sundelöf, A., Prista, N., and Fonseca, P. J. (**2020**). “Underwater noise recognition of marine vessels passages: Two case studies using hidden markov models,” ICES Journal of Marine Science **77**(6), 2157–2170.

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

Vieira, M., Fonseca, P. J., and Amorim, M. C. P. (**2021**b). “Fish sounds and boat noise are prominent soundscape contributors in an urban european estuary,” Marine Pollution Bulletin **172**, 112845.

Vieira, M., Fonseca, P. J., and Amorim, M. C. P. (**2024**). “Effect of boat noise on chorusing behavior of a marine fish (argyrosomus regius, sciaenidae),” in *The Effects of Noise on Aquatic Life: Principles and Practical Considerations* (Springer), pp. 869–876.

Vieira, M., Pereira, B. P., Pousão-Ferreira, P., Fonseca, P. J., and Amorim, M. C. P. (**2019**). “Seasonal variation of captive meagre acoustic signalling: a manual and automatic recognition approach,” Fishes **4**(02), 28.

Waddell, E. E., Rasmussen, J. H., and Širović, A. (**2021**). “Applying artificial intelligence methods to detect and classify fish calls from the northern gulf of mexico,” Journal of Marine Science and Engineering **9**(10), 1128.

Watson, R., Baste, I., Larigauderie, A., Leadley, P., Pascual, U., Baptiste, B., Demissew, S., Dziba, L., Erpul, G., Fazel, A. *et al.* (**2019**). "Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services," IPBES Secretariat: Bonn, Germany 22–47.

Yu, Y., Si, X., Hu, C., and Zhang, J. (**2019**). "A review of recurrent neural networks: Lstm cells and network architectures," Neural computation **31**(7), 1235–1270.

Zhang, H. (**2017**). "mixup: Beyond empirical risk minimization," arXiv preprint arXiv:1710.09412 .