

Article

Benchmarking an Integrated Deep Learning Pipeline for Robust Detection and Individual Counting of the Greater Caribbean Manatee

Fabricio Quirós-Corella ^{1,*} , Athena Rycyk ² , Beth Brady ³  and Priscilla Cubero-Pardo ⁴ 

¹ Advanced Computing Laboratory, National High Technology Center, National Council of Rectors, 98 Alexander Humboldt St., San José 10109, Costa Rica

² Pritzker Marine Biology, New College of Florida, 5800 Bay Shore Rd., Sarasota, FL 34243, USA; arycyk@ncf.edu

³ Save the Manatee Club, 317 Wekiva Springs Rd., Longwood, FL 32779, USA; brady@savethemanatee.org

⁴ National Council of Rectors, 98 Alexander Humboldt St., San José 10109, Costa Rica; pcubero@conare.ac.cr

* Correspondence: fquiros@cenat.ac.cr

Abstract

The Greater Caribbean manatee faces significant conservation challenges due to a lack of demographic data in low-visibility habitats. To address this, we present a refined automated manatee counting method pipeline integrating deep learning-based call detection with unsupervised individual counting. We resolved significant computational bottlenecks by implementing an offline feature extraction strategy, bypassing a 13-hour processing lag for 43,031 audio samples. To mitigate overfitting in imbalanced bioacoustic datasets, non-parametric bootstrap resampling was employed to generate 100,000 balanced spectrograms. Benchmarking revealed that transfer learning via a VGG-16 backbone achieved a mean 10-fold cross-validation accuracy of 98.92% ($\pm 0.08\%$) and an F1-score of 98.08% for genuine vocalizations. Following detection, individual counting utilized k-means clustering on prioritized music information retrieval descriptors—spectral bandwidth, centroid, and roll-off—to resolve distinct acoustic signatures. This framework identified three individuals with a silhouette coefficient of 79.20%, demonstrating superior cohesion over previous benchmarks. These results confirm the automatic manatee count method as a robust, scalable framework for generating the scientific evidence required for regional conservation policies.

Keywords: acoustic individual counting; bioacoustics; deep learning; feature extraction; kernel density estimation; machine learning; manatee call detection; music information retrieval; passive acoustic monitoring; transfer learning



Academic Editors: Sergio A. Velastin,
David Zabala-Blanco and Ruber
Hernández-García

Received: 20 December 2025

Revised: 14 February 2026

Accepted: 18 February 2026

Published: 3 March 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The sirenians, also known as sea cows, are a group of marine mammals whose populations have been declining annually. They are listed as threatened worldwide, with some subspecies experiencing regional functional extinction [1–5], such as the dugong populations (*Dugong dugon*) in China [6,7] and Japan [8]. A subspecies of the American manatee, the Greater Caribbean manatee (*Trichechus manatus manatus*) has a conservation status of vulnerable [9]. This herbivorous marine mammal is rarely observed off the Caribbean coast in Costa Rica, and its population status is unknown [1,10]. It has been affected by anthropogenic activities, like poaching, fishing, boat collisions, water pollution, and ecosystem degradation [1]. The Greater Caribbean manatee was declared a national symbol of

underwater fauna in Costa Rica in 2014, but the conservation regulations for this marine mammal are still scarce due to limited ecological knowledge around the species [1,10].

The 14th sustainable development goal (SDG), denoted as *Life Below Water* and developed by the United Nations (UN), emphasizes increasing scientific knowledge and building research capacity to support the conservation of marine biodiversity, which serves as an indicator of healthier oceans [11]. Hence, collecting and analyzing scientific evidence would justify establishing and implementing conservation actions to protect the manatee in coastal communities of Costa Rican Caribbean [1,12]. Traditional visual monitoring techniques—including shipborne, aerial, and land-based surveys—provide essential ecological data for sirenians but are often invasive and limited by adverse weather conditions, diurnal cycles, and prohibitive logistical costs [13–15]. Despite their substantial size, manatees are inherently difficult to track visually as they spend the vast majority of their life cycles submerged in shallow, turbid habitats where visibility is minimal [10,16]. In these environments, acoustic signaling serves as the primary modality for spatial navigation and conspecific communication [10,12,16,17].

The passive acoustics monitoring (PAM) serves as a critical strategy for assessing marine mammal abundance and spatio-temporal distribution, using the analysis of these bioacoustic signals without requiring direct visual observation [13,15,16,18]. This methodology offers a robust alternative to traditional visual surveys, particularly in marine environments characterized by high turbidity and limited light transmission where visual detectability is precluded [14,15,19,20]. Unlike invasive approaches such as satellite tagging, PAM is a non-disruptive, autonomous technique that minimizes human interference *in-situ* [20,21]. For instance, empirical studies have demonstrated that PAM yields higher success rates in detecting African manatees compared to sonar and visual point scans [20]. By deploying hydrophone arrays for durations ranging from several days to months, researchers can capture continuous acoustic data essential for characterizing diel activity and long-term phenological patterns in remote regions [13,14,18,20,21]. Because acoustic signals propagate significantly further underwater than light, PAM facilitates the detection of individuals at ranges extending several kilometers. This provides a scalable and cost-effective framework for monitoring populations while significantly increasing the likelihood of detecting target animals in inaccessible environments [13,18,20,21].

Bioacousticians differentiate individual manatees by analyzing specific acoustic features, most notably the (F_0), which demonstrates higher intra-individual stability than signal duration [22,23]. Supplemental diagnostic descriptors include spectral roughness [24], bandwidth (BW), and spectral roll-off [25]. By treating the vocalization's spectrogram as a unique acoustic “fingerprint,” analysts can resolve individual identities within a cohort [19]. Biologically, these vocalizations originate from laryngeal vocal fold vibrations and undergo modulation via the supralaryngeal vocal tract [26]. Typical signals range from 200 to 800 ms and exhibit significant ontogenetic variation; for instance, West Indian manatee (*Trichechus manatus*) calves produce notably shorter calls (89–420 ms) than adults [23,27,28]. While the F_0 generally spans 0.64 to 5.9 kHz, broadband components in wild populations may extend into ultrasonic frequencies reaching 150 kHz [23,27,28].

Vocalization rates in manatees are highly context-dependent, typically peaking during social interactions. Calves exhibit significantly higher signaling frequencies than adults, with production rates ranging from one call every 5 min to bursts of 27 calls/min in specific wild populations [22,28]. The resulting spectral characteristics reflect the interplay between the sound-producing laryngeal structures and the filtering properties of the vocal tract. Vibration of vocal fold homologs generates the F_0 [26], which varies systematically according to body size, age, and sex [22,23]. Younger calves frequently produce stereotypical “hill-shaped” frequency contours that tend to flatten as the animal matures and the

vocal tract lengthens [17]. These unique acoustic signatures facilitate essential individual recognition between lactating mothers and their offspring [22]. Although multiple individuals may vocalize simultaneously during high-arousal social interactions or mother-calf “duets”, the probability of complete temporal and spectral masking is mitigated by the distinctiveness of each individual’s harmonic structure [28,29].

As monitoring efforts scale toward year-round operations, the resulting longitudinal database (DB) renders labor-intensive manual annotation unfeasible, necessitating a transition to automated frameworks for manatee detection and individual identification [15]. While multi-feature representation spaces can resolve acoustically distinct groups within larger cohorts, the manual extraction and comparison of these parameters across massive DBs remain time-prohibitive, significantly constraining the scalability of traditional identification efforts [19]. Integrating artificial intelligence (AI) with high-performance computing (HPC) architectures enables the systematic analysis of long-term acoustic recordings in large DBs with minimal human supervision, ensuring that ecological insights into species presence and abundance remain both timely and statistically robust [13,18]. This research advances marine conservation by developing specialized, AI-driven applications tailored for the efficient interpretation of PAM data. By leveraging HPC resources, the proposed system facilitates high-confidence predictions of manatee presence and derived density estimations. These tools significantly enhance the ecological understanding of the Greater Caribbean manatee within regional waters, providing a scalable framework adaptable to monitoring other marine mammals across diverse aquatic soundscapes.

Related Work

The current literature reveals that traditional AI solutions for detecting and classifying single and multiple categories of marine mammals, especially cetaceans, rely on machine learning (ML) algorithms [13,18,30]. For example, humpback whale (*Megaptera novaeangliae*) acoustic detection has been achieved using the support vector machine (SVM) as a binary classifier to distinguish between song and non-song sounds in audio samples [13,18,30]. Furthermore, a hidden Markov model (HMM) has been implemented to identify both humpback whale calls and dolphin signature whistles, demonstrating its versatility in multi-species acoustic analysis [13,18,30]. Alternatively, an unsupervised approach employing a Gaussian mixture model (GMM) has proven effective in classifying vocalizations from various dolphin species and toothed whales, highlighting the potential of unsupervised ML in complex bioacoustic DBs [13,18,30]. Utilizing GMMs, a multi-classifier accurately has categorized vocalizations of Arctic marine mammals into seven classes: five whale species, walruses, and background noise, considering wavelets and Mel-frequency cepstral coefficients (MFCC) of the acoustic signals [31].

Recently, the deep ML or deep learning (DL) has shown an impressive throughput for marine mammals detection using different types of deep neuronal network (DNN) architectures [18,30]. A beluga whale (*Delphinapterus leucas*) classifier implemented an ensemble model of DNNs by training four distinct convolutional neuronal network (CNN) models, specifically designed for the detection of whistles and moans with the best accuracy score of 96.33% [32]. For the binary classification of humpback whale songs, a CNN framework incorporated active learning, where the outputs of candidate models were utilized to guide subsequent annotation efforts, thereby enhancing classification precision (97%) after training the models with these guided annotations [14]. Furthermore, a multi-classifier system for cetacean vocalizations was developed utilizing a seven-layer CNN architecture to categorize acoustic signals into six distinct whistle categories: constant, upsweep, downsweep, concave, convex, and sinusoidal [33]. This framework encompassed a broad taxonomic range, including 60 different marine mammal species in the DB to ensure robust

cross-species applicability. The model demonstrated significant reliability, achieving a mean classification accuracy of 95.20% across eight heterogeneous testing sets [33].

However, as DNNs scale in depth to enhance classification throughput by increasing model complexity, the gradients propagated during supervised training can diminish to near-zero values, compromising the learning at earlier layers [34]. To mitigate this vanishing gradients issue, residual learning bypasses layers and directly adds the input to the output through skip connections. The residual neuronal network (ResNet) architecture with dimensions of 512×256 leads to superior performance compared to CNNs on classifying 32 species of marine mammals by demonstrating a F1-score of 86.72% [34]. But it is crucial to acknowledge that ResNets, like many DL models, are susceptible to overfitting, particularly when they are trained on small datasets with several species categories or vocalizations types, as commonly occurs within PAM-DBs [30]. This means that there are not enough labeled audio samples to train these architectures due to the time-consuming nature of manual data labeling [18].

To mitigate the risk of overfitting and enhance generalization in complex DNN architectures within automated bioacoustic analysis, particularly when constrained by limited labeled DBs, *transfer learning* has emerged as a pivotal strategy. This DL technique utilizes pre-trained models that have been initially optimized on extensive, high-diversity datasets, allowing the architecture to leverage foundational feature hierarchies. By fine-tuning these models on a smaller, domain-specific corpus, the system inherits feature extraction (FE) capabilities while adapting to the unique spectral characteristics of bioacoustic signals. Recent applications have demonstrated the efficacy of this approach in multi-class marine mammal classification. For instance, researchers employed transfer learning using pre-trained CNN weights from ResNet-50 and VGG-19 to distinguish between three marine species, non-biological interference, and ambient noise, yielding accuracies of 95.30% and 96.10%, respectively [35]. Furthermore, specialized adaptations of AlexNet, utilizing replacement layers ranging from three to nine, have achieved a F1-score of 99.75% for signal detection and 98.50% for the multi-class categorization of killer whales, long-finned pilot whales, and harp seals [36]. These results underscore the capacity of transfer learning to provide high-fidelity detection and classification across diverse marine mammal taxa even when local training data is sparse.

Recent studies concerning sirenians bioacoustic analysis have predominantly leveraged ML/DL methods for detecting manatee vocalizations, utilizing various architectural approaches [25]. For instance, a binary classifier for Amazonian manatee (*Trichechus inunguis*) populations incorporated a CNN architecture, alongside an active learning methodology and data augmentation to address limited data, relying on manual inspection to increase labeled samples, resulting on an average precision of 98% for detecting manatee vocalizations in a training dataset [21]. To identify the Greater Caribbean manatee in the western Caribbean region of Panama, a data-driven scheme implemented binary classification for the using two CNN configurations: a linear architecture with a fixed kernel size and a pyramidal architecture with an increasing kernel size, including dropout, where all models achieved over 94% of accuracy for the testing test [37]. However, this specific system did not incorporate data augmentation or balancing techniques, potentially limiting generalization and increasing the risk of overfitting, which were actually suggested on this work, such the synthetic minority oversampling technique (SMOTE) to produce new data points [37].

For the binary classification of African manatee (*Trichechus senegalensis*) presence, a transfer learning framework was utilized to overcome the limitations associated with sparse labeled DBs and inherent class imbalance. This approach leveraged the pre-trained GoogLeNet architecture, which facilitates robust FE through its specialized inception

modules, even when trained on restricted volumes of domain-specific data [38]. While this implementation focused exclusively on signal detection rather than individual abundance estimation, it successfully mitigated the risks of model bias in high-noise underwater environments. The performance of the fine-tuned network was evaluated using a testing set from an entirely novel geographic location to assess its generalization capabilities. The model achieved a detection accuracy of 90.20%, demonstrating the efficacy of cross-domain knowledge transfer for identifying sirenians vocalizations in diverse ecological settings [38].

An alternative transfer learning framework utilized a dual-stage CNN ensemble for manatee vocalization classification, incorporating waveform-level data augmentation to improve model robustness and generalization [29]. This hierarchical architecture initially executes a binary classification to distinguish genuine manatee vocalizations from ambient noise, achieving a total accuracy of 91.15%. Subsequently, the system performs a five-class categorization—comprising squeaks, high squeaks, squeals, and mixed calls—utilizing linear spectrograms generated from waveforms resampled to 44.1 kHz, resulting in a classification accuracy of 92.86% [29]. EfficientNet served as the primary FE step, having been pre-trained on a zoo-based corpus and subsequently fine-tuned on longitudinal DBs from multiple captive environments to capture site-specific acoustic variations. Following successful classification, population size estimation was conducted exclusively using squeak calls to isolate unique individual signatures. The automatic FE stage for this unsupervised ML task utilized F_0 , specific frequency bands, and MFCCs as the core acoustic descriptors. To determine the census, high-density-based spatial clustering of applications with noise (HDBSCAN) was implemented as a divisive hierarchical clustering algorithm, paired with t-distributed stochastic neighbor embedding (t-SNE) for dimensionality reduction into a four-component feature space. When evaluated across various zoo settings, this counting framework demonstrated superior reliability, maintaining a lower population deviation (45.45%) than alternative unsupervised methodologies, thereby validating the efficacy of combining DL with density-based clustering for demographic monitoring [29].

In the turbid, low-visibility habitats of the Central American Caribbean and West Africa, manatees typically occur in localized groups of tens of individuals rather than thousands [19,20]. For example, in the San San River (Panama), population estimates using sonar have identified between 2 and 33 individuals depending on the season [19]. In Lake Ossa (Cameroon), a known manatee sanctuary, the population is estimated at approximately 49 individuals [20]. Therefore, an automated system in these regions is designed to resolve discrete acoustic signatures for a demographic scale of 10 to 100 animals in a given observation area [19,20]. Another approach corresponds to the fundamental framework for the identification and population estimation of the Greater Caribbean manatee on the Costa Rican Caribbean coast is established by the original automatic manatee count method (AMCM) pipeline. This computational architecture comprises a tripartite process: noise cancellation, vocalization detection, and individual recognition, integrating digital signal processing and unsupervised ML [24].

(AMCM) pipeline. This solution comprises a tripartite process based on signal processing and unsupervised ML: noise cancellation, vocalization detection, and individual recognition [24]. To enhance the detection range and sensitivity, the denoising stage employs an undecimated discrete wavelet transform (UDWT) and the the root mean square (RMS) value of the autocorrelation function. This computationally intensive phase incorporates resampling at 96 kHz and high-pass filtering at 2 kHz, followed by a moving average filter to attenuate artifacts due to snapping shrimps or water movement. Signal components are subsequently categorized using the k-means (KM) algorithm, which optimizes class separation based on maximum centroid distance to isolate biological signals from ambient noise [24]. Following signal enhancement, the call detection module utilizes a matched

filter approach integrated with the statistical characterization of the F_0 and peak frequency F_p of manatee vocalizations, applying a fixed threshold based on the median prominence of low-order harmonics to identify candidate vocalizations with a F-score measure of 96.7% at a SNR minimum of 10 dB [24]. The final individual recognition phase employs an FE and ML clustering routine. Three discriminative features—the mean-roughness logarithm, roughness standard deviation, and F_0 —were selected to characterize unique acoustic signatures. These features are processed using the expectation-maximization (EM) algorithm using a Gaussian mixture model (GMM) to cluster vocalizations detected, 4 manatees expected, ground truth, with the best results possible, exact number of individual and all calls correctly classified. To determine the optimal population count, a 10-fold cross-validation procedure iteratively increases the cluster count based on the average log-likelihood, terminating only when further expansion yields no statistical improvement in the model's predictive likelihood, [24]. To enhance the detection range and sensitivity, the denoising stage employs an undecimated discrete wavelet transform (UDWT) alongside the root mean square (RMS) value of the autocorrelation function (ACF). This computationally intensive phase incorporates signal resampling at 96kHz and high-pass filtering at 2 kHz, followed by a moving average filter designed to attenuate impulsive artifacts from snapping shrimp and hydrodynamic noise. Signal components are subsequently categorized using the k-means (KM) algorithm, which optimizes class separation via maximum centroid distance to effectively isolate biological sounds from the ambient acoustic background [24]. Following signal enhancement, the call detection utilizes a matched filter approach integrated with the statistical characterization of the F_0 and peak frequency F_p .

By applying a fixed threshold based on the median prominence of low-order harmonics, the system identifies candidate vocalizations, achieving an F1-score of 96.7% at a minimum minimum signal-to-noise (SNR) of 10 dB [24]. The final individual recognition phase employs an FE stage along with a clustering routine based on three discriminative acoustic features: mean-roughness logarithm, roughness standard deviation, and F_0 . These were processed using the expectation-maximization (EM) algorithm within a GMM framework to cluster detected vocalizations. To determine the optimal population count, a 10-fold cross-validation procedure iteratively evaluates the average log-likelihood; the cluster count is incremented until the model's predictive likelihood reaches a statistical asymptote, ensuring a perfect estimation of the expected number of individuals: 4 manatees for the corresponding testing set [24].

A comprehensive four-stage methodology to automatically detect and identify Antillean manatee vocalizations within the turbid wetlands of Panama [19]. The process begins with a detection stage that utilizes a five-level UDWT and ACF analysis to identify the harmonic structures typical of manatee calls while filtering out non-biological noise [19], replicating computation extensivity on the data processing stage on denoising and detection [24]. To improve signal quality, a denoising stage implements a signal-subspace approach using the Karhunen-Loève transform (KLT) to separate signal components from background interference, followed by a classification stage that employs a modified harmonic detection method to verify fundamental frequencies and harmonic presence [19].

Identification and counting stage, where vocalization spectrograms are treated as two-dimensional image patterns and represented using principal components analysis (PCA) coefficients, often referred to as “eigen-spectrograms”, which then feed into hierarchical and non-hierarchical clustering algorithms to distinguish individual manatees. Results from the Changuinola and San San rivers demonstrated high system sensitivity, with TP rates reaching 90–96% in the most effective classification modes [19]. In a scenario comprising 181 vocalizations grouped into ten operator-assigned acoustic clusters from field recordings, the clustering algorithm successfully organized the groups with a remarkably low error

rate of 4.42%. When applied to three years of field recordings, the methodology estimated population ranges of 33 to 34 likely individuals in the San San River and 45 to 48 likely individuals in the Changuinola River. By integrating the datasets from both river systems, the researchers identified a combined total of 63 to 68 likely unique individuals, a finding that supports the biological hypothesis of seasonal movement and habitat connectivity between these neighboring wetlands.

Experimental testing and performance profiling revealed that the denoising module of the original AMCM pipeline was computationally inefficient when processing long-term acoustic recordings and large-scale DBs. Consequently, subsequent research transitioned toward utilizing an HPC-powered approach centered on DNN methodologies [25]. This transition leveraged a transfer learning strategy employing VGG-16, which was fine-tuned on a labeled DB of Florida manatees (*Trichechus manatus latirostris*), achieving a binary classification accuracy of 96% [25]. To extend this framework toward population estimation, an unsupervised KM method, initialized via the EM algorithm, was implemented to group acoustic features. This hybrid approach effectively synthesized DL models for binary call detection with ML clustering to predict individual vocal origins, yielding a clustering score of 72% [25]. Despite these advancements, the system encountered significant challenges in real-world environments, primarily due to background noise interference from anthropogenic sources, such as boat engines and hydrodynamic splashing. Furthermore, performance was hindered by potential overfitting arising from inherent data imbalances and the restricted volume of the training corpus, highlighting the need for more robust regularization and data augmentation strategies in subsequent efforts [25].

(AMCM) framework proposed by Quirós-Corella et. al. (2024), focusing on enhancing binary classification robustness, pipeline scalability, and the precision of call clustering within complex marine soundscapes [25]. To achieve this, we introduce critical modifications to the time-frequency preprocessing stages and image data generation, while simultaneously refining prediction reliability measures and exploring advanced DL architectures. To mitigate overfitting and improve long-term inference stability, the methodology integrates strategic data balancing, bootstrapping-based augmentation, and rigorous cross-validation. These optimizations aim to provide accurate population and density estimations along the Costa Rican Caribbean coast. Ultimately, this research establishes a validated, high-performance framework designed to inform evidence-based conservation policies for environmental authorities and promote ecological awareness within coastal communities regarding the protection of the Greater Caribbean manatee. Therefore, this study addresses the limitations of the AMCM framework based on DL by enhancing binary classification robustness, pipeline scalability, and clustering precision within complex marine soundscapes [25]. The main contributions correspond to critical improvements to the time-frequency preprocessing and image generation stages, alongside enhance prediction reliability measures and the comparative evaluation between DL approaches. To mitigate overfitting and ensure long-term inference stability, the methodology integrates strategic data pipelining, balancing, bootstrapping-based augmentation, and rigorous cross-validation. These optimizations provide a standardized basis for accurate population and density estimations in the Costa Rican Caribbean coast. This research could establish a validated, high-performance framework designed to inform evidence-based conservation policies for environmental authorities. By improving the monitoring of the Greater Caribbean manatee, this system would promote ecological awareness and supports targeted protection efforts within coastal communities.

2. Materials and Methods

This study built upon the AMCM pipeline and was structured around two main steps: manatee call detection (MCD) and individual manatee counting (IMC) [25]. The MCD module implemented the necessary computational components for a binary classifier based on a comprehensive DL framework to discriminate between true and false manatee calls. Followed by the IMC module that employed ML clustering techniques to categorize detected calls based on music information retrieval (MIR) attributes for inferring acoustically possible individual manatees. The refined AMCM pipeline was designed for periodic offline analysis of accumulated PAM recordings. Typical data collection involves deploying one or more autonomous acoustic recording devices for weeks to months in diverse coastal habitats and locations, recording continuously on duty cycles [25]. Acoustic data are periodically retrieved and batch processed to generate temporal occurrence patterns of manatee calls and estimate the number of individual manatees that produced calls.

2.1. Training and Testing DBs

To train and validate the MCD model, a comprehensive dataset of acoustic recordings was constructed by integrating diverse PAM-DBs, totaling 43,031 labeled audio samples with variable signal duration and sampling rates ranging from 44.1 to 96 kHz. Florida manatee data included DTAG recordings (i.e., acoustic recording tags attached to a belt on the animal) from wild individuals in southwest Florida, land-based passive acoustic monitoring in Sarasota Bay, Florida, and underwater recordings from Blue Spring, Florida, and from Florida manatees under human care at ZooTampa [25]. African manatee recordings were acquired with underwater acoustic recorders in Lekki Lagoon, Nigeria, and Lake Ossa, Cameroon [39]. The inclusion of recordings from Florida manatees (*Trichechus manatus latirostris*) and African manatees (*Trichechus senegalensis*) was justified by structural similarities between these calls and those of the Greater Caribbean manatee [40]. The integrated DB consisted of WAV files binary labeled by a trained bioacoustician, categorized as *false vocalization* (28,765 negative entries) and *true vocalization* (14,266 positive instances), reflecting a class imbalance of 49.6%.

To assess the performance of the AMCM pipeline in its entirety under representative and similar field conditions of the Caribbean coast of Costa Rica, we selected 10 experimental recordings of the Greater Caribbean manatee from a DB from Bocas del Toro that consisted of samples with varying lengths and sampling rates [19,24]. The Panamanian and Costa Rican Caribbean regions are defined by their geographical proximity and shared narrow, sinuous riverine networks, such as the San San River and the Changuinola wetlands, which are separated by approximately 10km [19,25]. These habitats are characterized by brackish, highly turbid waters and dense aquatic vegetation, presenting a stark contrast to the clear-water spring refugia in Florida where other subspecies typically aggregate [16,17,19,25]. Such environmental disparities often induce significant domain shifts in DL architectures; models trained on high-quality acoustic data, such as recordings from Florida springs, frequently exhibit performance degradation when deployed in the sediment-heavy, low-SNR environments typical of Central American riverine systems [25,41].

Data acquisition for evaluating the AMCM pipeline involved hydrophone units permanently moored 15–20 m from the riverbank and 1 m above the river floor at depths of 2–3 m, facilitating continuous recording over several months [24,37]. These recordings remained unseen during the training phase of the MCD model, providing a rigorous benchmark for evaluating the detector generalization capabilities across novel acoustic environments [25]. Ground-truth validation was established through exhaustive manual inspection by trained bioacousticians who verified and tagged vocalizations to facilitate

individual identification and evaluate predictive accuracy [24]. While the testing PAM-DB provides a definitive number of expected vocalizations per recording, it lacks precise temporal indices for call onsets, necessitating a count-based validation approach for MCD inference [25]. The expected population count of four manatees within this corpus, based on visual inspection of spectrograms, serves as the primary benchmark for evaluating the population estimation performance of the IMC module [24].

Consequently, the output generated by the IMC module serves as a proxy count rather than an absolute census, facilitating an approximation of individual identification through the clustering of MIR attributes. This approach provides a statistical estimation of population density in scenarios where definitive individual labels are unavailable, establishing a robust metric for measuring relative abundance. This methodological constraint arises from the inherent limitations of the current experimental dataset, which catalogs the total number of expected vocalizations per sample without individual-specific ground-truth assignments. The testing DB also lacks spatial metadata and relational links between recordings; thus, it was utilized exclusively to evaluate the internal components of the IMC module. By leveraging the relationship between unique spectral signatures and discrete acoustic sources, the IMC framework establishes a scalable foundation for demographic analysis.

2.2. Computational Infrastructure

Experimental workflows for the MCD stage were executed on the Kabré HPC infrastructure, utilizing NVIDIA L40S-48GB and Tesla V100-PCIE-32GB GPU units to facilitate the training and evaluation of large-scale DNN architectures; comprehensive hardware specifications are detailed in Table 1. In contrast, the IMC module was executed exclusively on CPU resources, as the unsupervised clustering of the restricted feature space presented negligible computational overhead. This enabled a hybrid allocation strategy that prioritized GPU acceleration for high-dimensional classification tasks while offloading lower-complexity clustering operations to the host processor. To maximize computational throughput, mixed-precision training was implemented, leveraging 16-bit floating-point precision (float16) for compute-intensive tensor operations while maintaining numerical stability through 32-bit precision (float32) for critical variables, such as master weights and loss scaling. This approach effectively mitigated gradient underflow during the processing of augmented bioacoustic DBs. The NVIDIA L40S architecture demonstrated superior efficiency in executing both custom CNNs and pre-trained models without encountering memory leakage or allocation constraints during training or fine-tuning, establishing it as the optimal platform for this high-throughput supervised learning pipeline. Furthermore, comparative deployment testing between the V100 and L40S units during the model inference phase confirmed that while both architectures maintained functional parity.

Table 1. Detailed hardware configurations of the HPC nodes within the Kabré supercomputer facility. The table contrastively summarizes the heterogeneous environments used for benchmarking the AMCM pipeline, highlighting the evolution from legacy Tesla V100 systems to the high-bandwidth NVIDIA L40S architectures. These specifications establish the standardized computational baseline required for the execution time assessment of DNN training throughput and inference latency.

CPU	Memory	GPU	OS
2 × Intel Xeon Silver 4214R @ 2.40 GHz	31 GiB	1 × Tesla V100-PCIE-32GB	Linux 3.10.0-64bit
1 × Intel Xeon Silver 4416+ @ 2.00 GHz	256 GiB	1 × NVIDIA L40S-48GB	Linux 5.14.0-64bit

2.3. Feature Extraction and Pre-Processing

As illustrated in the architectural framework of Figure 1, the MCD module initiates with the WAV-DB generator, which manages the ingestion of storage directories contain-

ing either training-validation or testing datasets. This component utilizes a specialized file-fetching routine to interface with the HPC file system, systematically organizing raw acoustic datapaths into a standardized comma-separated value (CSV) transcription file. Depending on the operational mode, the functional block dynamically adapts its output to the experimental requirements. For supervised learning, it extracts and archives labeled samples; conversely, for inference in real-world environments, it generates a data structure of unlabeled WAV filenames. This systematic organization ensures consistent data flow into the subsequent processing stages, regardless of whether the system is undergoing model optimization or autonomous signal detection in unknown acoustic settings for the model trained.

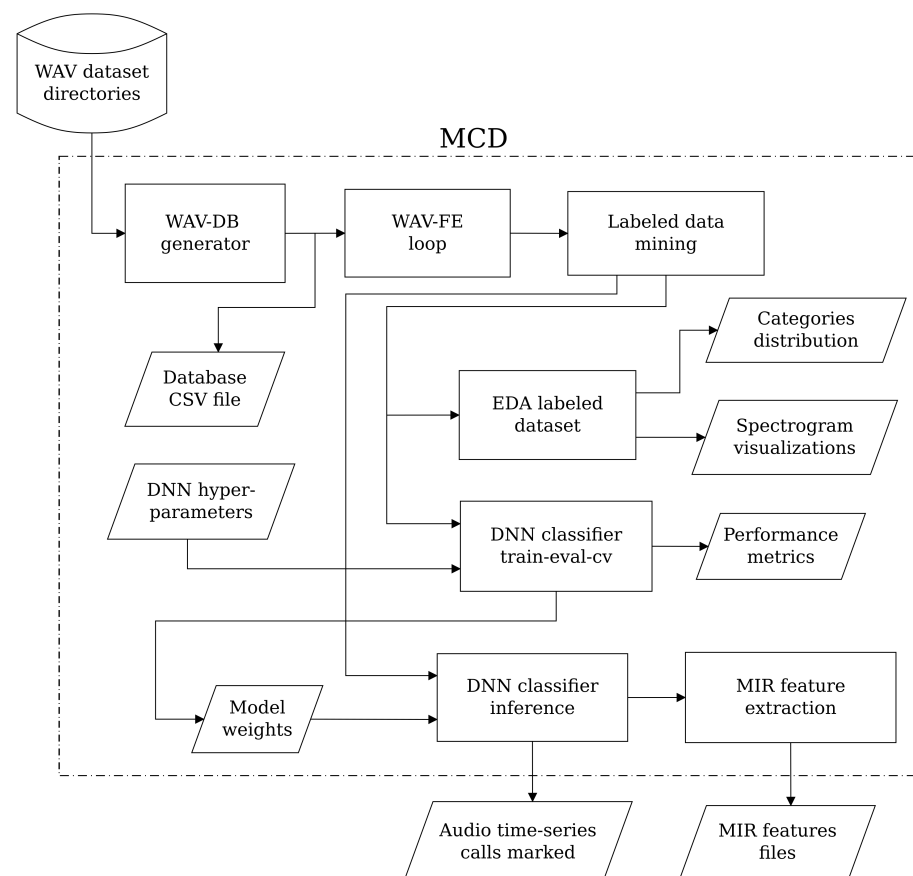


Figure 1. System architecture of the MCD module, depicting the dual-mode operational workflow for supervised training and autonomous inference. The schematic illustrates an integrated pipeline transitioning from raw acoustic data ingestion to high-fidelity FE with specialized preprocessing for signal enhancement. Key operational stages include bootstrap data augmentation integrated into the spectral mining process and an iterative DL model optimization utilizing cross-validation. This structural design ensures robust classification performance and methodological scalability within the broader AMCM framework.

Subsequently, the MCD module implemented a WAV-FE loop to transform raw 1-dimensional (1D) acoustic waveforms into discriminative 2-dimensional (2D) spectral representations through a standardized signal processing pipeline [18]. The core representation for the FE stage corresponded to the time-frequency data structure by applying successive fast Fourier transform (FFT) operations with a window size of 1022 and a 25% overlap, ensuring consistent spectral resolution across varying signal lengths. The transformation routine involved resampling input signals to a uniform 44.1 kHz and applying logarithmic interpolation to rescale spectral magnitudes, utilizing cutoff margins at 80% of the

frequency bounds. This process reassigns the linear frequency bins generated by the FFT onto a logarithmic scale, thereby enhancing the representation of the harmonic components and frequency modulations inherent in manatee vocalizations.

To enhance tonal clarity and isolate harmonic patterns within complex acoustic environments, a multi-stage denoising operation was implemented into the WAV-FE loop of Figure 1: (1) a 4th-order high-pass Butterworth filter with a 2 kHz cutoff frequency was applied to suppress low-frequency ambient interference and hydrodynamic noise; (2) spectral gating, configured for a 44.1 kHz sample rate, facilitated non-stationary noise reduction, utilizing a 1022-point FFT window and a 766-sample hop length to balance time-frequency resolution [42,43]; and (3) median filtering was employed for harmonic-percussive source separation (HPSS) to extract the horizontal spectral components, related to high-order harmonicals [44,45]. The HPSS configuration prioritized the isolation of tonal calls by implementing a harmonic filter width of 5 and a percussive filter width of 27. This was further refined by applying asymmetric margin sizes of (7, 1) for the harmonic and percussive masks, respectively. By emphasizing the vertical percussive suppression, the resulting spectrograms highlight the unique acoustic footprints of manatee vocalizations while mitigating impulsive interference. This integrated audio preprocessing ensures that subsequent MCD stages receive a high-SNR harmonic representation of the signal, which is critical for effective species detection.

dimensional (1D) waveforms into discriminative 2-dimensional (2D)-dimensional spectral representations [18]. Raw audio recordings were segmented into uniform 1-second temporal windows, selected to capture typical manatee call durations while minimizing call fragmentation across segment boundaries for inference mode, exclusively. Our study applied successive fast Fourier transform (FFT) operations with a window size of 1022 and 25% overlap to achieve time-frequency conversion. The core pipeline involved resampling signals to a uniform 44.1 kHz and applying logarithmic interpolation to rescale spectral magnitudes, with cutoff margins set at 80% of the upper and lower frequency bounds. This reassigned the linear frequency bins generated by the FFT onto a logarithmic scale, which better represented the harmonic structure of manatee calls. dimensional (1D) waveforms into discriminative 2D spectral representations [18]. The WAV-FE loop of Figure 1 applied successive fast Fourier transform (FFT) operations for each labeled audio sample with a window size of 1022 and a 25% overlap to achieve time-frequency conversion. The core pipeline involved resampling signals to a uniform 44.1 kHz and applying logarithmic interpolation to rescale spectral magnitudes, with cutoff margins set at 80% of the upper and lower frequency boundary values. The operation effectively reassigned the linear frequency bins generated by the FFT onto a logarithmic scale, which better approximated the perceived harmonic structure of biological signals.

Traditional FE methodologies relied on onset detection and padding to spot patterns of interest by enforcing uniform segment lengths within unconstrained acoustic data prior data preprocessing and FE transformation [25]. In contrast, this study implemented a comprehensive image processing stage that directly generates 2D spectrograms on a dB-scale, transforming audio into a visual format optimized for DL models. These representations were further processed with logarithmic interpolat, horizontal flipping, and pixel-interpolation resizing before being converted into $128 \times 128 \times 3$ RGB color images via color-space conversion (CSC). This workflow encapsulated acoustic signatures into discrete spectral images stored in 8-bit unsigned integer (uint8) format to ensure architectural compatibility and standardized pixel intensity ranges.

(EDA) was conducted to assess spectral image quality and verify that preprocessing maintained distinguishable acoustic characteristics between classes. (EDA) was included to assess spectral image quality and verify that preprocessing maintained distinguishable

acoustic characteristics between true and false vocalizations by plotting random samples from the complete JPEG-DB. To address the computational bottleneck inherent in the WAV-FE loop depicted in Figure 1, an image preprocessing step was implemented as an alternative FE variant for labeled DBs. This approach involved the asynchronous generation and storage of spectral representations as JPEG files during the WAV-FE loop, indexed alongside their corresponding ground-truth labels. By pre-calculating these transformations, the execution overhead during the benchmarking of diverse DL architectures was significantly reduced, allowing for more rapid iteration cycles. The subsequent FE for image data loaded the structured spectrograms to yield a high-dimensional tensor of $43,031 \times 128 \times 128 \times 3$, representing the cumulative time-frequency data of the training and validation splits. To maintain rigorous data integrity, a centralized CSV transcription file was generated, mapping each unique image path to its respective class.

2.4. Data Mining and Augmentation

The data mining phase illustrated in Figure 1 involved the vectorization of the pre-processed spectral data into a high-dimensional array of $43,031 \times 49,152$. Following the verification and removal of not-a-number (NaN) values to ensure dataset integrity, explicit external scaling techniques—such as min-max normalization or standardization—were bypassed. This design choice was justified by the integration of batch normalization layers within the DL architectures under consideration, which effectively regulated input distributions and mitigated internal covariate shift during training. To ensure architectural compatibility with the binary classification loss functions, raw target classes were mapped into a one-hot encoded representation. This transformation resulted in a labels array of $43,031 \times 2$, where [1. 0.] and [0. 1.] denotes false and true vocalization, respectively. To maximize computational throughput for GPU-accelerated processing, both feature vectors and class labels were cast to the float32 data type. This configuration balanced memory management with spectral resolution, maintaining numerical consistency and high efficiency throughout the training and inference workflows.

Following the data mining phase illustrated in Figure 1, the labeled DB was partitioned into three distinct subsets—training (70%), testing (20%), and validation (10%)—to facilitate a rigorous assessment of model generalization. A stratified random sampling approach was employed to strictly maintain the proportional representation of each class across all partitions, ensuring a consistent baseline for the comparative analysis of the binary classifier. This strategy aims to preserve the underlying data distribution within each subset, thereby mitigating variance in performance metrics and ensuring that the evaluation remains representative of the global dataset. Specifically, the training set comprised 30,121 samples (20,135 false and 9986 true vocalizations), the testing set contained 8607 spectrograms (5754 negative and 2853 positive), and the validation set included 4303 images (2876 negative and 1427 positive). This distribution maintains an inherent class imbalance of approximately 50% across the MCD module, accurately reflecting the skewed nature of real-world acoustic data where biological signals are typically sparse.

To mitigate the inherent class imbalance of the training corpus and enhance model generalization, non-parametric bootstrap resampling with replacement was applied independently to each partition, precluding data leakage between subsets [46]. This iterative process involved the random selection of spectrograms, where each sample was returned to the pool to allow for potential duplication. Specifically, the training minority ($n = 9986$) and majority ($n = 20,135$) classes were resampled to achieve a balanced target of 35,000 samples per category, increasing the training set from 30,121 to 70,000 observations. A symmetric procedure was executed for the validation partition, where the minority ($n = 1427$) and majority ($n = 2876$) classes were resampled to a uniform distribution of 15,000 samples per

category, expanding the set from 4303 to 30,000 observations. In contrast, bootstrapping was omitted for the test set ($n = 8607$) to avoid data duplicates leakage and to ensure evaluation on genuine samples that reflect the natural imbalance between false and true vocalizations. This augmentation strategy yielded a consolidated dataset of 100,000 labeled spectrograms for training and validation, which were vectorized into a multidimensional float32 array to optimize computational throughput and numerical consistency for DL integration.

(EDA) throughout the data mining phase to derive valuable insights regarding the effects of audio preprocessing on both types of spectral image samples [25]. This included analyzing data distribution across categories, dimensions, data type, and visualizing the samples transformed in the WAV-FE module to observe the highlighted patterns within the images associated with the tonal calls of the target species. A data pipelining architecture was implemented to mitigate input-output bottlenecks due to high-dimensional arrays and prevent hardware idling. This framework optimized memory utilization through integrated operations, including stochastic shuffling, automated batching, and asynchronous pre-fetching to ensure a continuous data stream for internal layers. All balanced subsets were integrated into float32 data pipelines, decoupling computationally intensive preprocessing from model execution to maximize throughput during large-scale training and inference. As illustrated in Figure 1, exploratory data analysis (EDA) was conducted following the data mining phase to quantify the impact of audio preprocessing on spectral characteristics. This analytical process involved visually inspection of original and replicated samples processed from data mining phase. Such visualization was critical for validating the preservation of distinctive harmonic patterns and frequency modulations within the tonal calls of the target species. By verifying the content and integrity of the data pipeline results, this step ensured that the spectral images accurately captured the diagnostic acoustic patterns required for effective binary classification.

2.5. Model Building and Configuration

The proposed MCD implementation was developed within a supervised DNN framework—integrating functional blocks for training, validation, evaluation, cross-validation, and inference—ensuring methodological continuity with prior research [25]. To identify the optimal architectural paradigm, we benchmarked a custom-designed CNN model, recognized for their efficacy in marine bioacoustics [18], against transfer learning approaches utilizing pre-trained VGG-16 and VGG-19 backbones. While additional architectures—including custom ResNets, ResNet-50, ResNet-101, EfficientNet-B3, and MobileNet-V2—were evaluated, they exhibited higher rates of overfitting compared to the custom CNN and provided no performance gains; consequently, we reported in this manuscript the benchmarking of the top-three DL architectures that we found during the MCD implementation with improvements on WAV-FE and data mining blocks in Figure 1. The adoption of transfer learning specifically addresses the data scarcity challenges inherent in bioacoustic monitoring, a strategy that has demonstrated significant effectiveness in recent sirenians vocalization studies [16,25,29].

All the DL architectures built were standardized with an input shape of $128 \times 128 \times 3$ to accommodate color spectrogram dimensions and an output layer of two artificial units to align with the binary one-hot encoding scheme. The Adam optimizer was implemented with a baseline learning rate of 1×10^{-5} , while numerical stability and overfitting were addressed through L_2 regularization ($\lambda = 0.001$) and a 0.7 dropout rate applied to the final layers. The models utilized ReLU activation for hidden layers and a Softmax function for the terminal probability distribution, ensuring a robust mapping of acoustic features to class likelihoods. Hyperparameter optimization was executed through a systematic brute-force search for both DL paradigms, evaluating architectural configurations across a 4×3

dimensional hyperparameter space. This iterative process involved testing varying CNN layer depths (2, 4, 6), kernel dimensions (3, 5, 7), filter counts (8, 16, 32), and dense units (32, 64, 128) within the classification head. Performance was rigorously assessed against training-validation and cross-validation metrics, with final architectural selections informed by comparative analysis against previous experimental configurations.

The custom-designed and configurable architecture for the CNN model was comprised of four hidden layers, each layer sequentially followed by max-pooling and batch normalization layers. The CNN layers employed a spatial dimension of 3×3 and initiated with 16 filters, which subsequently doubled per layer (i.e., $16 \rightarrow 32 \rightarrow 64 \rightarrow 128$). The final classification block consisted of a dense, fully-connected, layer with 64 artificial neurons, feeding into the final binary output layer; the model construction and configuration of the custom CNN model resulted in a network with 106,306 trainable parameters. And for transfer learning, the model ensemble was constructed upon either a pre-trained VGG-16, serving as the baseline from previous work [25], or VGG-19 architecture, both initialized with weights trained using ImageNet dataset [47]. The pre-trained convolutional base was augmented with a custom classification head, consisting of a single fully-connected layer with 128 units. The initial backbone involved 65,922 trainable parameters for both pre-trained VGG-16 and VGG-19 architectures. Subsequently, the fine-tuning stage involved unfreezing all base layers, resulting in a larger parameter count for optimization: approximately ~ 7.15 M parameters for VGG-16 and ~ 9.50 M parameters for VGG-19.

2.6. Model Training, Evaluation, and Cross-Validation

Model training for each DL methodology was standardized at 600 epochs with a batch size of 128, utilizing binary accuracy and cross-entropy as the objective functions to quantify the divergence between predicted and actual class distributions. To prevent overfitting and optimize computational expenditure, the training workflow integrated early stopping and model checkpointing. The early stopping mechanism was configured to monitor validation loss with a minimum improvement threshold of 0.0001 and a patience of 10 epochs. Simultaneously, the model checkpointing routine archived the optimal state in a .keras file format whenever a performance improvement was detected, ensuring that only the most generalizable weights were preserved for subsequent inference.

Regarding the transfer learning strategy, it was executed through a two-stage sequential optimization: an initial FE stage and a global fine-tuning phase. In the primary phase, the base layers were frozen while 25% of the terminal CNN layers were optimized at a learning rate of 1×10^{-5} . Subsequently, the entire architectural ensemble was unfrozen for fine-tuning at a reduced learning rate of 1×10^{-6} , facilitating high-fidelity adaptation to the specific spectral signatures of the manatee vocalizations. Following the convergence of the fine-tuning stage, the model was restored to its optimal parameter set to discard weights from the terminal, potentially overfitted epochs. The validation process concluded with an EDA suite, generating historical training-validation progress curves and visual classification outputs to verify the diagnostic efficacy of the MCD module.

Upon completion of the DL training and validation phases, the MCD module implemented model evaluation using an independent subset of labeled 2D spectrograms to assess generalization performance on unseen samples [25]. Model inferences are categorized into four fundamental outcomes: true positive (TP), representing a correct identification of a manatee vocalization; false positive (FP), denoting an incorrect detection of a vocalization signal; true negative (TN), indicating a correct identification of ambient noise; and false negative (FN), signifying a missed detection of a non-vocalization signal. These outcomes formed the basis of the confusion matrix, a diagnostic tool utilized to quantify the trade-offs between detection sensitivity and false alarm rates within complex marine environments.

negative (FN), signifying incorrect non-positive class predictions. With these metrics allowed the computation of a classical tool for binary and multiple, classification, the confusion matrix, summarizes the classifier performance. To rigorously characterize the classifier's performance, a comprehensive suite of metrics was derived from these classification outcomes [25]. Accuracy defined in Equation (1), provided a global measure of predictive correctness across both true and false vocalizations:

$$\text{accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}}. \quad (1)$$

Precision and recall (sensitivity) were calculated to evaluate detection purity and completeness for both categories, as defined in Equations (2) and (3), respectively.

$$\text{precision} = \begin{cases} \text{true vocalization} : \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{false vocalization} : \frac{\text{TN}}{\text{TN} + \text{FN}} \end{cases}, \quad (2)$$

$$\text{recall} = \begin{cases} \text{true vocalization} : \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{false vocalization} : \frac{\text{TN}}{\text{TN} + \text{FP}} \end{cases}. \quad (3)$$

The F1-score of Equation (4), representing the harmonic mean of precision and recall, served as a robust indicator of model performance under potentially imbalanced conditions:

$$\text{f1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (4)$$

Finally, the model's discriminative power was quantified also through the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. By evaluating the TP rate against the FP rate across varying decision thresholds, this metric provided a threshold-independent measure of the classifier's ability to distinguish biological signals from ambient noise.

To rigorously assess generalization and mitigate overfitting, a k -fold cross-validation strategy was implemented with $k = 10$ mutually exclusive partitions. The dataset was divided into 10 equal-sized subgroups, following a protocol where the model underwent 10 training iterations; in each cycle, $k - 1$ folds were utilized for supervised learning, while the remaining held-out fold served as the unseen validation set [30]. This rotating validation framework ensures that every sample is utilized for testing exactly once, effectively neutralizing potential biases associated with static data partitioning. By calculating the error on each excluded fold and aggregating the results, the system obtained unbiased estimates of both training and validation errors. This comprehensive methodology provides a more robust performance metric than a single split, ensuring that the reported accuracy and loss reflect the model's true capacity to generalize across novel bioacoustic datasets [30].

2.7. Model Inference and MIR-FE

During the model inference stage, the WAV-FE pipeline was adapted to process continuous, unlabeled field recordings while maintaining architectural parity with the training configuration [25]. The inference sequence initiates with the ingestion of raw acoustic data, followed by resampling and temporal segmentation into uniform 1 s windows. These intervals were strategically selected to encapsulate typical manatee vocalization durations

while minimizing signal fragmentation across segment boundaries. To handle simultaneous vocalizations, the MCD module leverages the characteristic short-duration nature of manatee calls (200–800 ms) [12]. Following segmentation, each waveform window undergoes the identical preprocessing and FE sequence established during the training phase to ensure structural consistency within the spectral data. This systematic approach preserves localized signal integrity through the vectorization of multidimensional spectrogram arrays and the casting of features to float32 precision. By utilizing optimized data pipelining, the system achieves high-throughput processing, enabling the efficient transformation of raw bioacoustic recordings into standardized input tensors compatible with the deployed DL architectures.

Following the completion of the WAV-FE and data mining phases, the system instantiated the DL model and loaded optimized weights to initiate the inference process. Predictions were generated by iteratively processing data batches through a specialized pipeline tailored to each experimental recording in the DNN inference block of Figure 1. During this routine, indexed spectrogram arrays were fed into the architecture to obtain class-specific posterior probabilities, which were subsequently archived in a comprehensive transcription CSV file for downstream analysis. To refine the results and mitigate FP occurrences, a decision threshold was applied to the prediction confidence scores, serving as a critical filter to distinguish valid manatee vocalizations from ambient environmental noise during inference. Specifically, only detections exceeding a probability threshold of 0.5—the central tendency of the distribution—were classified as genuine calls. While this parameter remains configurable, values below this limit were discarded as non-vocal high-energy noise. This criterion ensures that the resulting census and spatio-temporal density estimations are derived exclusively from high-confidence acoustic events, thereby preserving the integrity of the subsequent IMC module.

As a fundamental component of the EDA for the inference stage and final output of Figure 1, a 1D-signal representation was generated for each input sample. This time-domain visualization was annotated with temporal markers indexing the specific instances where the model identified potential manatee vocalizations [25]. To augment this analysis, the system implemented a random sampling routine to visualize batches of spectral images, each labeled with its segment index, predicted class, and associated confidence score. This visual framework facilitated qualitative assessment of model performance on unseen data, providing essential insights into generalization capabilities within complex marine acoustic environments. Furthermore, this dual-representation approach enabled the visual verification of TPs and the identification of FPs, thereby streamlining the process for subsequent expert validation and iterative model refinement [25].

To validate the behavioral performance of the MCD module on unseen data, inference outcomes were evaluated using an experimental repository of real-world, labeled field recordings [24]. This verification process leveraged apriori knowledge regarding the expected number of manatee calls within each recording to benchmark the model's predictive accuracy and quantify its generalization capabilities. The inference performance was measured using a modified experimental error metric that quantifies the deviation of model estimations from the expected values:

$$\text{inference error} = \left| \frac{\text{predicted} - \text{expected}}{\text{expected}} \right| \quad (5)$$

where *predicted* denotes the number of calls estimated by the model and *expected* represents the ground-truth vocalization count, generated by bioacoustic experts that provided the testing DB [25].

To establish a robust quantitative foundation for individual recognition, a comprehensive bioacoustic analysis was executed to extract specialized MIR descriptors of each detected vocalization in the testing DB utilized on the MCD module. This procedure constitutes the foundational FE stage for the subsequent population estimation within the IMC module. Initially, an adaptive time-domain denoising routine after resampling (44.1 kHz) was applied to each identified call in the input signal, integrating high-pass filtering, signal normalization, and spectral gating. This preprocessing pipeline was specifically engineered to isolate harmonic components of the vocalizations detected and emphasize the F_0 structure, thereby mitigating the influence of non-stationary ambient noise and enhancing signal-to-noise ratios.

For each enhanced signal, 11 distinct MIR attributes were computed to characterize the unique acoustic signatures of the vocalizations processed. These descriptors, encompassing F_0 , BW, spectral centroid, spectral roll-off, spectral contrast, spectral flatness, zero-crossing rate (ZCR), kurtosis, skewness, and the RMS value, form the high-dimensional feature set derived from the MCD module output [25]. Upon completing the inference and MIR-FE iterations across the testing DB, the system generated a structured data array mapping detected calls against their respective acoustic features. This consolidated feature space facilitates the unsupervised clustering routines within the IMC module, enabling approximated population and spatio-temporal density estimations.

2.8. Unsupervised Learning

with noise (HDBSCAN), and agglomerative clustering. Consistent with the IMC methodology established in prior research, the preparation of unlabeled samples via the MIR-FE pipeline remained fundamentally uniform, as depicted in Figure 2 [25]. The current study utilized a structured dataset that integrated manatee vocalization observations with a suite of descriptors spanning both the temporal and spectral domains. From this framework, the IMC module extracted a specialized MIR data array for vectorization. A rigorous data mining phase preceded the clustering analysis, involving a systematic inspection to identify and remove observations containing NaN entries. These invalid feature values were frequently associated with residual FP detections that bypassed the initial MCD inference stage. By purging these artifacts, the system ensures that the subsequent unsupervised ML routines are grounded in valid acoustic signatures of genuine vocalizations, thereby enhancing the reliability of the approximated individual identification and population density estimations.

Following data cleaning, we applied feature scaling via min-max normalization to constrain the feature values within the range of 0 to 1, ensuring feature value parity and preventing descriptors with larger numerical scales from dominating the model training. To further refine the feature space and optimize the subsequent unsupervised ML modeling, a multi-stage dimensionality reduction strategy was implemented. Initially, feature selection was performed by calculating the Gini importance derived from fitting a random forest (RF) classifier with 100 decision trees (DT) to the MIR feature map [25]. This process facilitated the identification and isolation of the top-three most relevant acoustic descriptors, effectively reducing noise and computational complexity while preserving the most discriminative information for the clustering algorithm.

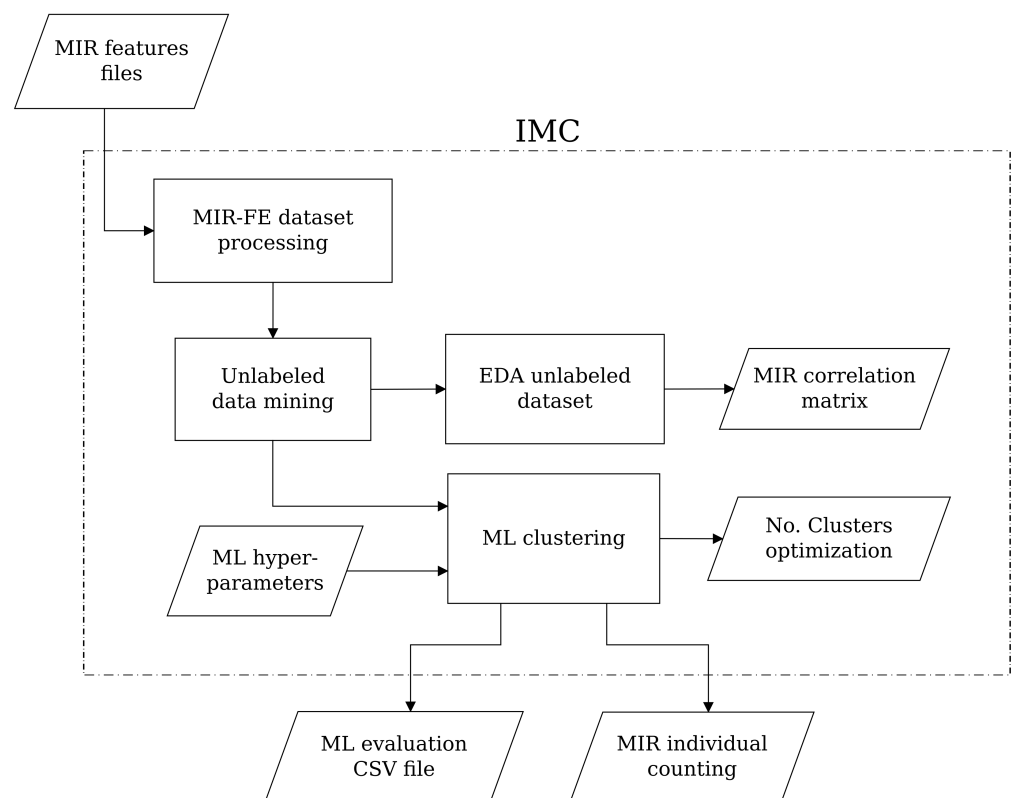


Figure 2. Architectural overview of the IMC module within the AMCM framework, illustrating the computational pipeline for population assessment. The diagram delineates the integration of MIR dataset generation and data mining with unsupervised ML clustering to establish a proxy for individual counting. This modular workflow facilitates the characterization of identified vocalizations in a set of data and the estimation of manatee abundance, bridging PAM recordings with biological population metrics.

To optimize the MIR feature space, PCA coefficients were utilized to compress the multidimensional acoustic descriptors into a two-dimensional representation [25]. These dimensionality reduction techniques were implemented to mitigate the curse of dimensionality, thereby refining the input space for unsupervised ML and population estimation. This reduction facilitates more efficient convergence of algorithms by emphasizing the most significant variance within the bioacoustic data while filtering redundant information. Methodological consistency was maintained by replicating the clustering selection strategy and empirical tuning parameters from previous benchmarks [24,25]. This alignment with both the MCD development and established research frameworks ensures a rigorous comparative analysis of the system’s ability to estimate population density from acoustic patterns.

As part of the EDA step for the IMC module (Figure 2), a correlation matrix was calculated to quantify relationships between MIR attributes, effectively identifying latent dependencies and data redundancies. Complementary bivariate scatter plots were utilized to visualize the distribution of potential individual manatees across paired MIR descriptors, providing a graphical assessment of feature patterns derived from real-world acoustic DBs [25]. To enhance output interpretability of the IMC module with EDA, bivariate scatter plots as kernel density estimation (KDE) visualized the distribution of potential individual manatees across paired MIR descriptors.

The unsupervised IMC module of Figure 2 implemented the EM-KM methodology [24,25], utilizing Lloyd’s algorithm to facilitate rapid convergence. To ensure deterministic

reproducibility, centroid initialization was performed using a seeding method, which samples initial points based on their empirical probability distribution contribution to overall inertia. The parameter selection for the EM-KM method was optimized through a systematic trial-and-error approach. The model was standardized with a maximum of 400 iterations per run, 100 independent initializations, and a convergence tolerance of 0.0001 relative to the within-cluster sum of squares.

Model performance of the IMC model was validated through a dual-metric approach incorporating inertia and the silhouette score. The optimal cluster count k , representing distinct individuals with unique MIR descriptors, was determined via the elbow method by identifying the inflection point of the inertia curve. Moreover, the silhouette score (Equation (6)) was employed to quantify cluster cohesion and separation, where $a(i)$ denotes the mean intra-cluster distance and $b(i)$ the mean distance to the nearest neighboring cluster:

$$\text{silhouette score} = \frac{1}{N} \sum_{i=1}^N \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (6)$$

To finalize the IMC outputs, the structured transcription CSV files were updated, archiving vocalizations with their respective MIR attributes and cluster assignments to support downstream ecological modeling [25]. The MIR transcription produce a certain of acoustic signature, so double-counting is prevented by the temporal stability of these signatures; research indicates that an individual's vocal parameters remain consistent across recording dates and even over several years [19].

3. Results

3.1. Acoustic Data Processing

Processing the complete labeled DB via the WAV-FE pipeline for training the MCD model required a total execution time of 12 h, 4 min, and 10 s utilizing the NVIDIA L40S-48GB GPU unit. This significant execution time overhead underscored a major computational bottleneck within the supervised training workflow of the MCD module (Figure 1). To address this challenge and accelerate the benchmarking of various DL architectures, a decoupled, defined as an offline or asynchronous strategy for the WAV-FE step was implemented. By pre-generating and archiving the spectral images through an isolated execution of the WAV-FE loop, the system eliminated redundant signal processing during model training. This approach significantly enhanced the efficiency of the experimental framework, facilitating rapid architectural iteration and hyperparameter tuning without the recurrent cost of raw waveform transformation.

As a foundational EDA component of the WAV-FE pipeline, Figure 3 illustrates a representative data batch, after completing the MCD data mining phase. This data batch consists of randomly selected spectrograms rendered as RGB images, employing an ordinal encoding scheme for ground-truth binary labels, where 0 denotes a *false vocalization* and 1 signifies a *true vocalization*. Within this visualization, pixel intensities range from 0 to 255, as indicated by the associated color scale. Because these samples are presented for initial EDA rather than active AMCM inference, they represent raw input data without associated model predictions or confidence scores. The visualization of these curated subsets facilitates the identification of diagnostic acoustic patterns characteristic of the target species. Valid vocalization samples reveal distinct high-frequency harmonic patterns—frequently exceeding 2–4 kHz—and vertical frequency modulations that contrast sharply against the stochastic nature of ambient underwater noise. This qualitative inspection verifies that the sequential WAV-FE pipeline, JPEG compression, and data pipelining stages successfully preserve the spectral integrity required for robust DL classification.

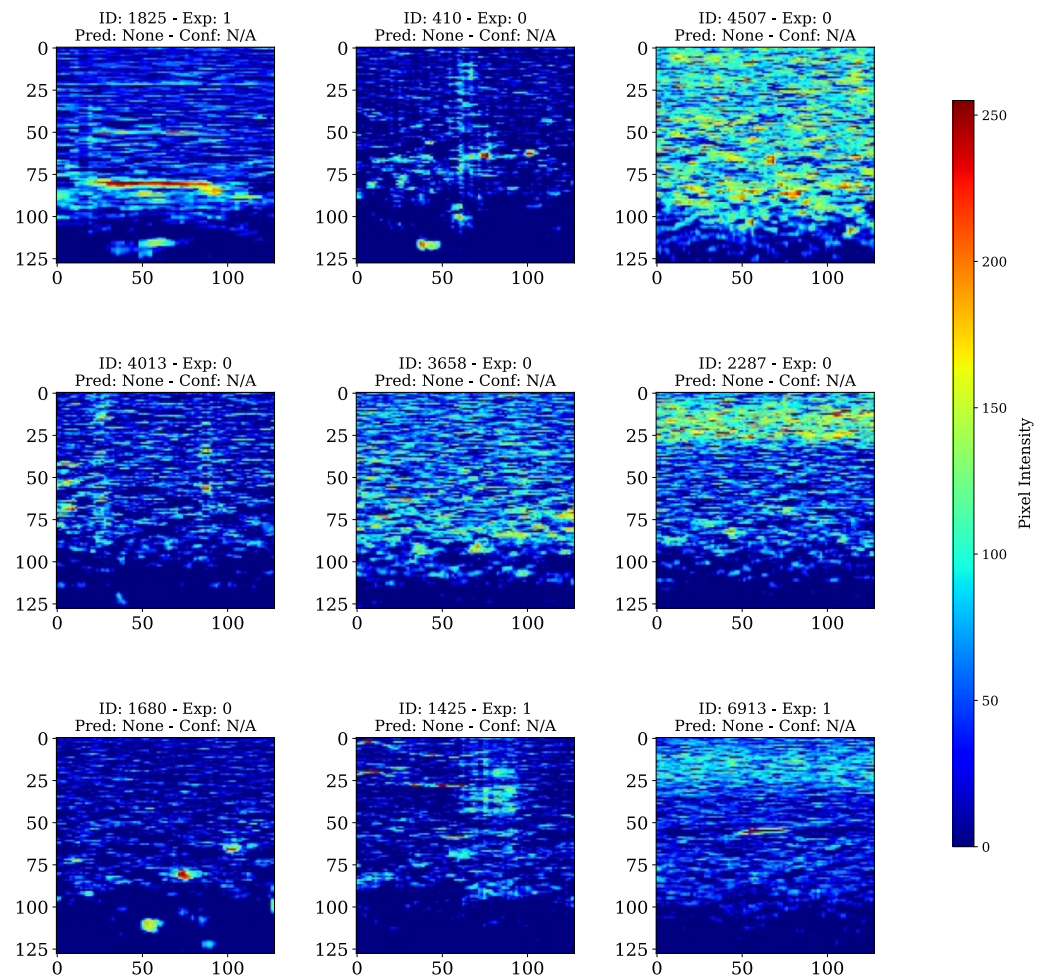


Figure 3. EDA visual suite of a representative spectrogram batch sampled from the evaluation pipeline after data augmentation. The visualization displays the indexed spectral images generated via offline FE and archived as JPEG files, where binary labels 0 and 1 denote false and true vocalizations, respectively. This diagnostic verification confirms the successful preservation of harmonic structures and acoustic signatures within the standardized input tensor, ensuring data integrity.

3.2. Model Benchmarking

Considering that EDA debug was enabled, storing intermediate outputs for all the experiments performed, supervised training of the custom CNN architecture was completed in 29 min and 55.67 s on the NVIDIA L40S GPU, with the early stopping mechanism terminating optimization at the 98th epoch, as illustrated in Figure 4a,b. The model checkpointing callback preserved the optimal weights identified at the 88th epoch, exporting the model in .keras format for subsequent deployment. For the transfer learning paradigms, the initial FE phase with frozen weights lasted 30 min and 20.81 s for VGG-16 and 1 h, 12 min, and 32.05 s for VGG-19. Subsequent fine-tuning achieved convergence more rapidly; VGG-16 required 17 min and 11.22 s, as depicted in Figure 5a,b, while VGG-19 concluded in 11 min and 31.44 s (Figure 5c,d). Total backpropagation durations and finalized MCD modeling execution times are summarized in Table 2. Due to its larger parameter space, VGG-19 exhibited significantly longer training durations than the VGG-16 pretrained weights and the custom CNN architecture. Figure 5 delineates the transition between the frozen-layer phase and global optimization for each pre-trained model; VGG-16 initiated fine-tuning at the 103rd epoch and converged at the 156th, whereas VGG-19 transitioned at the 234th epoch and reached terminal convergence by the 267th. In all instances, checkpointing ensured the retention of the highest-performing weights discovered during the fine-tuning stage for further evaluation.

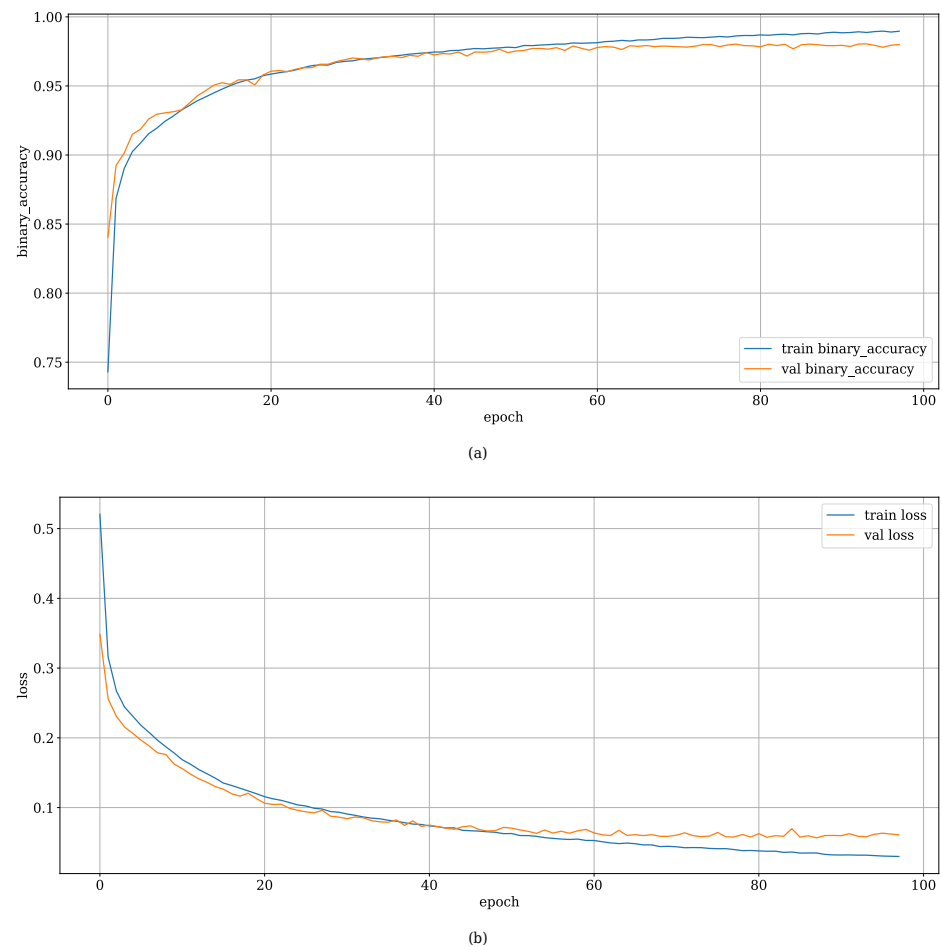


Figure 4. Learning curves of the custom-designed CNN architecture during supervised optimization. The plots illustrate the longitudinal progression of (a) binary accuracy and (b) binary cross-entropy loss across training and validation subsets, with termination indicating the point of optimal weight preservation via early stopping.

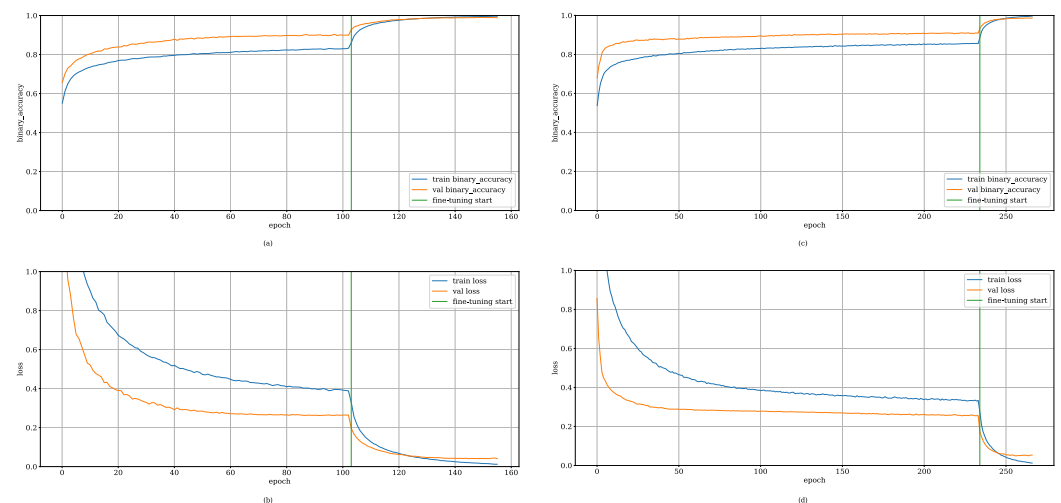


Figure 5. Comparative analysis of transfer learning dynamics for VGG-16 and VGG-19 backbones. Subplots (a,b) represent VGG-16 metrics, while (c,d) summarize VGG-19 performance. The trajectories demarcate the dual-fitting strategy, highlighting the efficacy of early stopping in maximizing generalization on the manatee bioacoustic dataset.

The optimized DL models were reconstructed to assess generalization performance on an independent evaluation split, quantifying classification throughput and error distribution across both target categories. As summarized in Table 2, these metrics underscore

the high statistical stability inherent across all paradigms. The VGG-16 backbone, in particular, provided the optimal balance between predictive sensitivity and categorical error minimization for manatee vocalization estimation. Furthermore, detailed classification performance is elucidated via the confusion matrices in Figure 6.

Table 2. Comparative training execution time and performance metrics for supervised MCD architectures evaluated on the independent testing subset. The summary evaluates the custom-designed CNN against pre-trained VGG-16 and VGG-19 backbones using global binary accuracy, binary cross-entropy loss, and AUC-ROC. These results justify the selection of the VGG-16 architecture as the study’s baseline [25], as it demonstrates superior error minimization and discriminative stability over both the custom CNN and the VGG-19 variant.

Model	Duration [s]	Accuracy	Loss	AUC-ROC
CNN	1.80×10^3	98.01%	6.05%	97.97%
VGG-16	2.85×10^3	98.72%	4.58%	98.59%
VGG-19	5.04×10^3	98.46%	5.32%	98.33%

		Predicted (CNN)	
		false vocalization	true vocalization
Actual	false vocalization	98.12%	1.88%
	true vocalization	2.17%	97.83%

		Predicted (VGG-16)	
		false vocalization	true vocalization
Actual	false vocalization	98.97%	1.02%
	true vocalization	1.79%	98.21%

		Predicted (VGG-19)	
		false vocalization	true vocalization
Actual	false vocalization	98.66%	1.34%
	true vocalization	2.00%	98.00%

Figure 6. Normalized confusion matrices evaluating binary classification performance for the benchmarked architectures. The matrices quantify predictive precision and recall for the MCD module in discriminating between manatee presence (*true vocalization*) and environmental noise (*false vocalization*). These results corroborate the selection of the VGG-16 backbone as the optimal baseline [25], demonstrating superior sensitivity and specificity compared to both the custom CNN and the VGG-19 variant.

The custom-designed CNN correctly identified 5646 TNs and 2791 TPs, incurring 108 FNs and 62 FPs. In contrast, the VGG-16 classifier demonstrated superior error suppression, recording 5695 TNs and 2802 TPs, with only 59 FNs and 51 FPs. The VGG-19 variant yielded comparable reliability, identifying 5677 TNs and 2796 TPs alongside 77 FNs and 57 FPs. The normalized results underscore a robust discriminative capacity across all evaluated architectures, particularly in the accurate detection of manatee presence under established conditions. While the custom CNN maintained competitive performance, the transfer-learning backbones—led by the VGG-16 architecture—consistently yielded the lowest marginal error rates. The performance metrics summarized in Table 3 corroborate the high reliability of these benchmarked models. Specifically, the custom-designed CNN achieved F1-scores of 98.52% and 97.04% for negative and positive samples, respectively. Transfer learning variants provided incremental performance gains; notably, the VGG-16 backbone yielded F1-scores of 99.04% and 98.08% for false and true vocalizations, while the VGG-19 variant produced F1-scores of 98.83% and 97.66%.

Table 3. Detailed performance breakdown of the MCD module across the testing subset, categorized by architectural paradigm and classification target. The report summarizes precision, recall, and F1-score for identifying manatee calls and distinguishing them from environmental noise. The VGG-16 model emerges as the superior configuration, achieving the highest harmonic mean between predictive sensitivity and positive predictive value.

Model	Class	Precision	Recall	F1-Score
CNN	false vocalization	98.12%	98.91%	98.52%
	true vocalization	97.83%	96.28%	97.04%
VGG-16	false vocalization	98.98%	99.11%	99.04%
	true vocalization	98.21%	97.94%	98.08%
VGG-19	false vocalization	98.66%	99.01%	98.83%
	true vocalization	98.00%	97.32%	97.66%

To qualitatively evaluate performance on unseen data, Figure 7 presents an EDA visual suite of VGG-16 inference outputs using the non-bootstrapped testing partition. Figure 7a illustrates correct predictions with confidence scores approaching 100% for both true and false vocalizations, confirming the architecture’s capacity to maintain high-confidence detections across varying SNR levels. These results underscore the model’s robustness in identifying biological signals amidst diverse background noise. In contrast, Figure 7b highlights misclassifications where confidence scores are markedly lower than those of correct predictions. These errors primarily emerge from short-duration calls or vocalizations obscured by environmental artifacts. For instance, the classifier appears to confuse harmonic components of false vocalizations, such as in spectrogram ID:1555 with a confidence of 97.07%. Despite the preprocessing applied during the FE stage prior model testing, high-interference samples like ID:4621 can mask positive signals with high confidence (99.38%), leading to FNs.

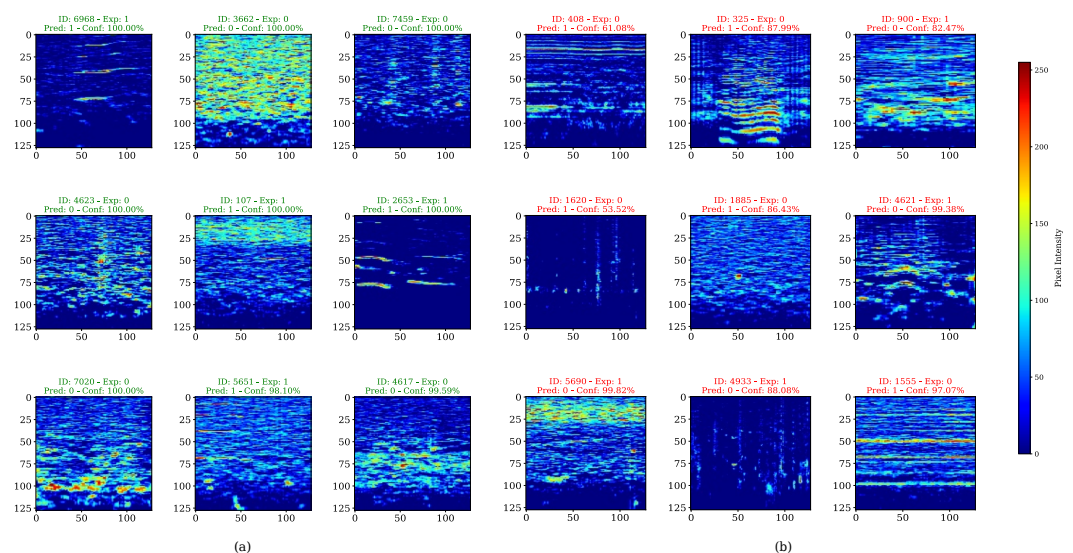


Figure 7. Visual validation of VGG-16 inference on a representative testing batch. Each subplot displays the input spectral representation alongside the corresponding prediction probability and categorical assignment. Green annotations in subplot (a) denote successful alignment between model predictions and expert-annotated ground truth, while red annotations in subplot (b) identify misclassifications. This qualitative diagnostic confirms the model’s capacity to extract discriminative harmonic features from complex acoustic environments, maintaining high-confidence classifications across varying SNR levels.

A 10-fold cross-validation routine was executed on the NVIDIA L40S unit to rigorously assess the statistical stability and generalization capacity of the benchmarked models,

as summarized in Table 4. The custom-designed CNN completed the procedure in 7 h, 30 min, and 2.848 s, yielding a mean accuracy of 97.93% with a mean binary cross-entropy of 5.78%. While a linear extrapolation from a single 30-minute training session might suggest a shorter baseline, the observed duration accounts for the cumulative computational overhead inherent in iterative data re-partitioning, model re-initialization, and post-fold metric aggregation. The transfer learning architectures demonstrated superior predictive performance, albeit at the cost of increased computational requirements. The VGG-16 backbone emerged as the most robust configuration, maintaining a mean binary accuracy of 98.92% and a mean loss of 4.06% over a total execution time of 9 h, 45 min, and 26.79 s. In comparison, the VGG-19 variant exhibited slightly lower aggregated performance, with a mean accuracy of 98.55% and a loss of 4.94% over 12 h, 33 min, and 44.66 s. The minimal standard deviations recorded across all paradigms in Table 4 confirm the models' high discriminative stability, regardless of specific data partitions.

Table 4. Temporal profiling and execution latency of the critical MCD inference stages on the NVIDIA L40S-48GB GPU unit. It quantifies the computational overhead for heterogeneous field recordings at varying sampling rates (F_s), contrasting raw signal length (**Duration**) against the online WAV-FE stage and MCD inference. The results identify signal FE as the primary throughput bottleneck, while the core DL inference maintains high-speed processing regardless of the initial sampling frequency.

Sample	F_s [kHz]	Duration [s]	WAV-FE [s]	MCD [s]
S(1)	96	20.37	15.98	1.35
S(2)	48	18.71	12.14	0.13
S(3)	96	34.08	21.71	0.18
S(4)	96	7.39	5.07	0.12
S(5)	96	17.36	12.35	0.13
S(6)	48	13.07	9.77	0.12
S(7)	48	16.35	10.58	0.13
S(8)	48	3.78	2.56	0.12
S(9)	96	59.36	38.26	0.17
S(10)	96	57.32	36.72	0.18

3.3. Manatee Call Detection

Following our model benchmarking, the MCD system transitioned to the deployment phase for inference on unseen acoustic recordings. The VGG-16 transfer learning ensemble was selected as the core architecture, a decision supported by its superior performance during the empirical evaluation and cross-validation stages based on multiple classification metrics. Compared to the custom CNN and VGG-19 configurations, the VGG-16 backbone consistently yielded higher cross-validation accuracy and demonstrated a more robust balance between architectural depth and generalization capacity. This optimal configuration effectively addressed the specific acoustic constraints of the manatee bioacoustic DB, outperforming alternative models in classification stability. The complete system was deployed for experimental inference on an unseen testing DB comprising 10 recordings from the Bocas del Toro data repository [25].

Total execution time—comprising model instantiation, online spectrogram generation via WAV-FE, DNN inference, and MIR evaluation with active EDA debugging—totaled 24 min and 35 s on an NVIDIA L40S-48GB GPU. This represents a significant performance acceleration over the legacy Tesla V100-PCIE-32GB, which required 33 min and 55 s. Table 4 provides a granular breakdown of execution latency per recording for critical pipeline stages during MCD inference on the high-performance L40S architecture. The testing DB represented a total signal duration of 247.79 s ($\bar{x} = 24.78$ s), with sample lengths ranging from 3.78 to 59.36 s at sampling rates of 48 kHz and 96 kHz. Analysis reveals that the WAV-

FE loop constituted the primary temporal overhead, requiring 164.33 s ($\bar{x} = 16.43$ s). In contrast, the core MCD inference was significantly more efficient, totaling only 2.64 s ($\bar{x} = 0.26$ s), with a peak execution duration of 1.31 s during the initial inference pass—a latency primarily attributed to the one-time model initialization overhead. This discrepancy confirms that the computational bottleneck resides in the WAV-FE stage rather than DNN classification. These results highlight the high inference efficiency of the optimized architectures for real-time bioacoustic monitoring.

The inference performance of the MCD module, summarized in Table 5, contrasts expert ground-truth annotations with model predictions to quantify predictive variance per sample through the normalized error metric of Equation (5). Quantitative evaluation across the external dataset yielded 100 total predictions against a baseline of 69 expected vocalizations, resulting in an aggregate inference error of 0.45. While the system achieved a high detection rate, the module exhibited an average per-recording error of 0.62. These metrics suggest that high sensitivity to acoustic events may lead to predictive variance, potentially driven by environmental noise interference or call over-segmentation. Notably, significant discrepancies in samples S(3) and S(7) highlight the module’s capacity to identify low-amplitude vocalizations previously omitted during manual labeling, as well as marginal FPs triggered by transient acoustic artifacts in the Caribbean environment.

Table 5. Inference performance benchmarking of the MCD module on the NVIDIA L40S-48GB GPU unit across unseen field recordings from the Bocas del Toro DB [25]. The table contrasts manually annotated ground-truth (**Expected**) against MCD-generated candidates (**Predicted**) and MIR-validated detections (**Valid**). High error rates observed in samples S(3) and S(7) illustrate the module’s sensitivity to low-amplitude calls—frequently omitted during manual labeling—while identifying marginal FPs triggered by high-noise events.

Sample	Expected	Predicted	Valid	Error
S(1)	9	10	9	0.11
S(2)	6	7	7	0.17
S(3)	6	17	16	1.83
S(4)	5	5	5	0.00
S(5)	4	7	7	0.75
S(6)	4	5	4	0.25
S(7)	3	9	3	2.00
S(8)	2	1	1	0.50
S(9)	16	22	22	0.38
S(10)	14	17	16	0.21

Manual analysis of sample $S(1)$; for instance, demonstrated high precision with a marginal inference error of 0.11. The model generated 10 candidates against an expected count of 9; qualitative evaluation via EDA transcription confirmed the additional detection possessed a confidence score of 0.62. However, manual validation determined the signal lacked the characteristic harmonic structure of a manatee vocalization, identifying it as a FP. This artifact was subsequently flagged during the MIR data-cleaning phase of the IMC module and removed as a NaN value, resulting in 9 valid detections as reported in Table 5. In sample $S(2)$; for example, the system achieved an inference error of 0.17, identifying seven detection markers (CALL:1–7) across the 19 s signal. This discrepancy primarily resulted from temporal over-segmentation, where detections CALL:3 and CALL:4 captured a single continuous vocalization spanning two consecutive 1 s windows. As Figure 8 revealed, all detections reached classification probabilities exceeding 0.99, correlating high-confidence decisions with distinct spectral signatures. Figure 8a illustrates the raw waveform of the $S(2)$ recording, providing high-density calls in a 19 s excerpt to visually demonstrate model performance. Complementing this, Figure 8b presents a random selection of spectrograms from the corresponding sample, displaying class assignments and confidence scores to exemplify manatee presence estimation in an unseen recording. Such segments at different sampling rates are atypical of standard PAM deployments, which generally exhibit highly imbalanced distributions where valid vocalizations represent only 1–5% of the total recording duration.

Conversely, sample $S(3)$ exhibited the maximum recorded inference error of 1.833 in Table 5. Post-analysis inspection of the spectral outputs revealed that the majority of predicted signals were TPs previously overlooked during manual ground-truth annotation. This finding suggests that the MCD module possesses a higher sensitivity for subtle, low-amplitude signals than the initial human baseline, effectively identifying valid acoustic events within complex underwater environments. As a final stage, the MCD module executed the MIR-FE routine for every identified vocalization in each testing sample, extracting temporal and spectral descriptors from the acoustic signals to generate 10 individual MIR transcription files.

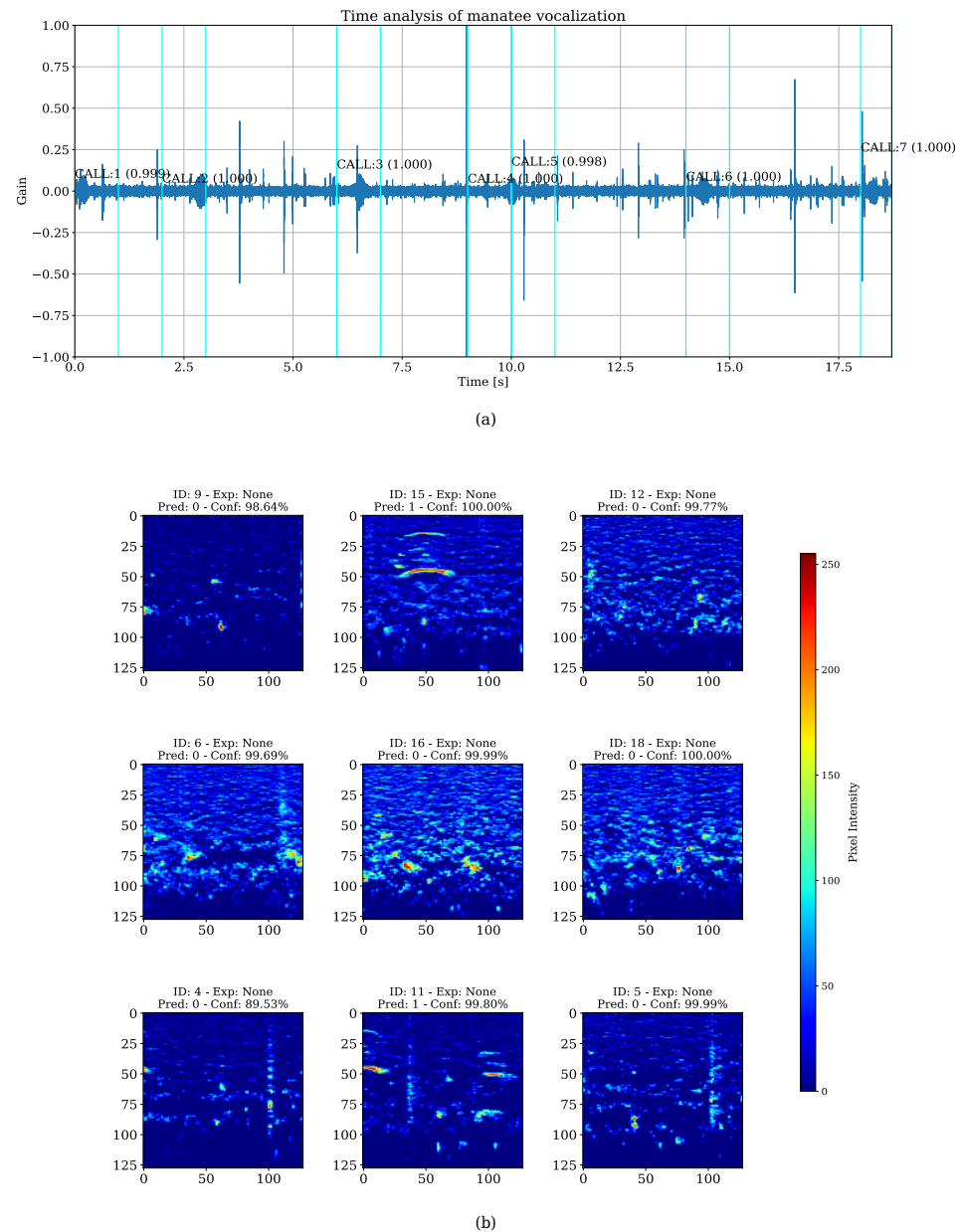


Figure 8. Visual representation of the MCD inference process applied to unseen field recording $S(2)$ using the VGG-16 ensemble. Subplot (a) illustrates the time-domain 1D waveform for a detected segment, featuring the estimation markers and its prediction probability. Subplot (b) presents the corresponding batch of 2D spectral images utilized as inputs for the DNN architecture. This qualitative assessment verifies the architecture's proficiency in identifying discriminative harmonic signatures within complex ambient noise.

3.4. Individual Manatee Count

forest (RF) classifier using the estimated labels by KM algorithm, or the dimensionality reduction (e.g., PCA, NMF). As EDA results, Pearson correlation results of the input data array, size of reduced dataset. The MIR transcriptions generated by the MCD module were consolidated via the IMC module into a unified, unlabeled acoustic dataset, as illustrated in Figure 2. This initial feature matrix comprised 100 observations, each characterized by 11 acoustic and temporal descriptors. During the IMC data-cleaning phase, 10 NaN values associated with identified FPs were removed, resulting in a refined dataset of 90 valid TPs, with the distribution per sample summarized in Table 5. To optimize computational efficiency, the cleaned data were vectorized and cast to float32 precision. Subsequently, min-max normalization was applied to scale all features within the $[0, 1]$ range. This prepro-

cessing step is critical for distance-based clustering analysis, as it ensures that descriptors with disparate physical units—such as frequency in Hz and duration in ms—contribute proportionally to the similarity metrics.

To improve the subsequent unsupervised ML phase, feature selection was performed utilizing Gini importance with RF models, identifying spectral BW, spectral centroid, and spectral roll-off as the primary descriptors. This selection resulted in an 90×3 feature matrix. Subsequent EDA revealed that these descriptors exhibited high positive Pearson correlation coefficients, ranging from 0.95 to 0.99 (Figure 9), indicating significant redundancy within the feature space. To mitigate multicollinearity and provide a de-correlated input space, PCA was applied to reduce the dimensionality from three descriptors to two principal components. This transformation yielded a finalized 90×2 data structure that preserved the underlying signal variance while facilitating more distinct cluster separation. The optimal number of clusters was determined to be $k = 3$ by iteratively evaluating the inertia score across a range of all potential counts using the elbow method.

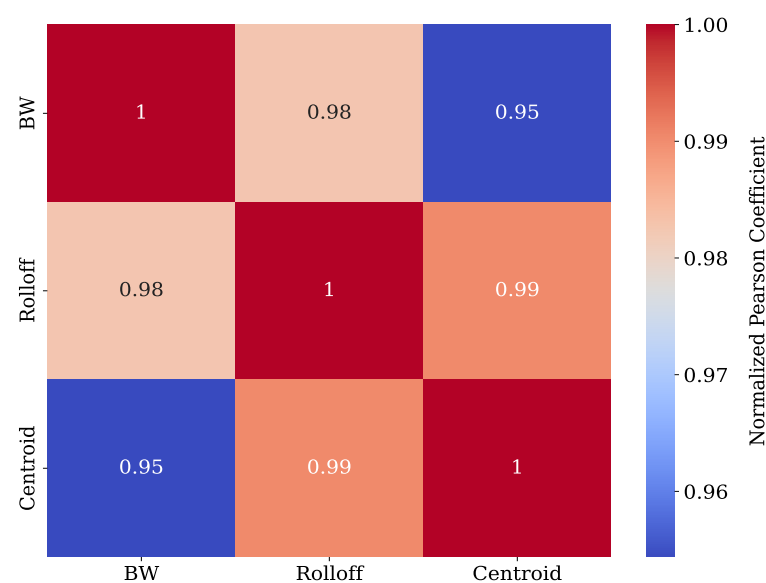


Figure 9. Pearson correlation analysis of the optimized acoustic feature set utilized for IMC clustering. The matrix illustrates the linear relationships between spectral descriptors derived from the MIR-FE process, specifically highlighting the top three descriptors prioritized by RF Gini importance. High coefficients observed among these primary features justified the subsequent application of PCA for dimensionality reduction.

forest (RF) classifier. The efficacy of the IMC clustering was rigorously validated without supervision using the silhouette score (Equation (6)), where the EM-KM framework achieved a robust coefficient of 79.20%. This result indicates significant intra-cluster cohesion and distinct inter-cluster separation, confirming that the MIR feature space provides a reliable basis for individual acoustics differentiation. Following this validation, cluster assignments were integrated into the MIR dataset to enable multi-parametric analysis. As illustrated in Figure 10, bivariate analysis of F_0 and BW approximated a characterization of spectral distributions via KDE, establishing a quantitative foundation for the spatio-temporal estimation of manatee populations. The model successfully resolved three distinct MIR groups, each characterized by a unique vocal signature: *Individual 1* (red, $n = 21$), *Individual 2* (yellow, $n = 57$), and *Individual 3* (green, $n = 12$). Spectral profiling identified that *Individual 2* possessed the narrowest BW (1.0–1.75 kHz) and the highest F_0 range (≈ 10 kHz), whereas *Individual 3* exhibited a broader BW peaking near 2.25 kHz with a lower F_0 . *Individual 1* was characterized by a specific BW range between 2.50 kHz and 2.75 kHz. For

the IMC evaluation, the system predicted three individuals against an expected count of four [24], yielding an inference error of 0.25 according to Equation (5).

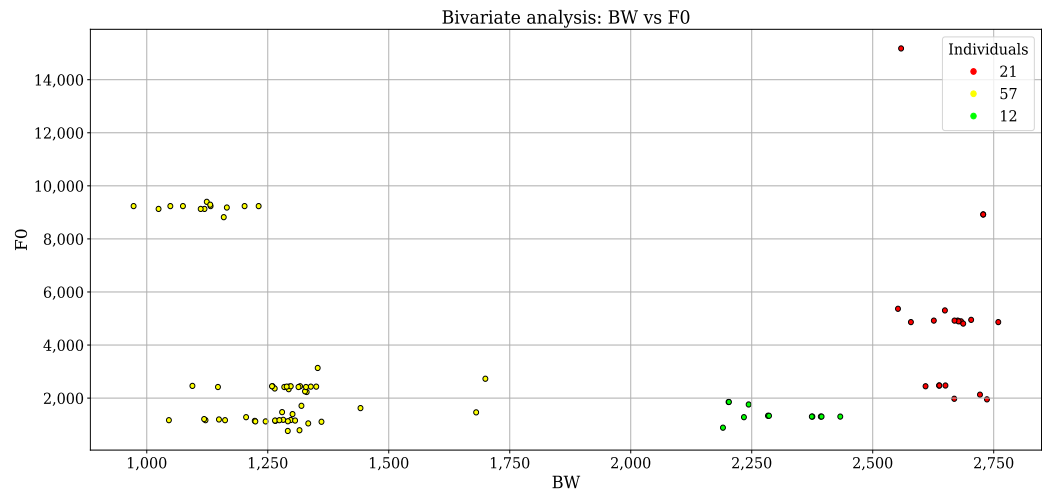


Figure 10. Bivariate density analysis of spectral descriptors for unsupervised manatee population estimation. The KDE visualization illustrates the distribution of detected vocalizations within a feature space defined by the F_0 and spectral BW. Clusters identified through the IMC framework represent distinct acoustic signatures, serving as a biological proxy for individual identification. This spatial partitioning provides a quantitative framework for differentiating population subgroups and estimating the abundance of vocalizing individuals within a set of field recordings.

4. Discussion

The implementation of a decoupled, offline WAV-FE strategy represents a critical resolution to the computational bottlenecks identified in prior sirenians bioacoustic research [24,25]. This DL framework for the MCD module contrasts with earlier AMCM iterations that relied on computationally intensive UDWT transformations at 96 kHz for signal denoising and KM clustering for evaluating call candidates, which frequently induced hardware idling and memory segmentation faults [24]. Transforming 43,031 variable-length audio samples into $128 \times 128 \times 3$ RGB spectrograms initially required nearly 13 h of continuous GPU processing. By decoupling the WAV-FE loop from the training phase and pre-storing spectral images as JPEG files, the MCD pipeline (Figure 1) eliminated redundant signal processing during architectural benchmarking, a significant improvement over previous frameworks [25]. Furthermore, the transition from legacy Tesla V100 systems of previous work to high-bandwidth NVIDIA L40S architectures, as detailed in Table 1, facilitated the processing of large multidimensional arrays generated by intensive data augmentation and pipelining. This migration significantly reduced latency during fine-tuning and optimized memory management for large-scale pre-trained weights. Findings confirm that leveraging modern GPU architectures for batch processing enables the rapid transformation of raw WAV data into standardized spectral images, providing a sustainable and scalable framework for conservation actions. This optimization ensures high-throughput analysis across extensive regional bioacoustic DBs, collected via PAM methods.

A primary challenge in developing DL architectures for manatee bioacoustics is the inherent scarcity and imbalance of labeled DBs, which significantly elevates the risk of overfitting. The initial curated dataset exhibited a substantial 50% imbalance between target categories. To mitigate this and enhance model generalization, a non-parametric bootstrapping method was applied post-data splitting, yielding a balanced corpus of 100,000 spectrograms for training and validation. This methodology addresses critical limitations in previous Caribbean manatee classifiers that lacked robust augmentation or balancing strategies [25,37]. However, the selected method introduces a notable trade-

off: the random replication of existing samples may restrict the model's exposure to the natural stochastic variability of underwater acoustics, potentially challenging pattern recognition during real-world PAM inference. Despite these constraints, the refined WAV-FE pipeline—incorporating resampling, high-pass filtering, spectral gating, HPSS, and logarithmic frequency interpolation—successfully framed diagnostic visual patterns associated with manatee tonal information. Furthermore, transitioning from traditional windowing to image resizing via pixel interpolation allowed for the encapsulation of complete acoustic information from samples of varying durations into standardized labeled RGB spectrograms. This processing suite sharpened the discriminative boundary between valid calls, ambient background noise, and visually similar negative artifacts. EDA visualizations in Figure 3 confirm these results, revealing high-frequency harmonic structures and well-defined spectral features that facilitate a more accurate MCD stage. Nevertheless, the aggressive nature of the pre-processing may occasionally produce invasive artifacts or additive noise in certain negative samples, potentially impacting inference sensitivity. Complementary to mixed-precision support, optimized data pipelining proved essential for efficient memory management following the augmentation phase, ultimately enabling high-throughput deployment within the HPC infrastructure.

Although extensive automated hyperparameter optimization was initially considered, a manual trial-and-error configuration was adopted to prioritize computational efficiency. This decision was grounded in the rationale that a carefully tuned architecture is naturally more resilient to overfitting, whereas exhaustive brute-force searches demand powerful HPC resources. This strategic allocation ensured that available computation was concentrated on signal processing enhancement and architectural benchmarking rather than redundant hyperparameters optimization cycles. Experimental analysis of the MCD module focused on a rigorous benchmarking of task-specific supervised learning against high-capacity transfer learning architectures. This comparison facilitates a performance evaluation between personalized models and pre-trained backbones adapted via fine-tuning, ultimately identifying the most robust solution for manatee acoustic identification. The custom CNN, VGG-16, and VGG-19 were selected as the primary reporting models due to their effectiveness in achieving high performance despite the data scarcity inherent in marine mammal classification. Prior to this selection, a comprehensive exploration of alternative architectures, including custom ResNets and pre-trained weights of ResNet-50, ResNet-101, EfficientNet (B3), and MobileNet (V2), was conducted. These models did not yield results superior nor an acceptable fitting, compared to the selected paradigms, thereby validating the architectural choices presented in this study. While the custom CNN offered superior execution throughput due to its lower parameter count, the VGG-16 architecture—though computationally more intensive—demonstrated superior evaluation metrics and a more stable accuracy convergence following terminal fine-tuning.

The classification metrics summarized in Tables 2, 3 and 6 establish a performance benchmark for the MCD modeling, the most critical stage of the AMCM framework. The custom-designed CNN, optimized with regularization and dropout, achieved a robust mean 10-fold cross-validation accuracy of 97.93% ($\pm 0.08\%$) and a testing AUC-ROC of 97.97%, exceeding previous transfer learning benchmarks [25]. Despite this performance, the refined transfer learning approach—leveraging enhanced time-frequency FE and a balanced bootstrapping strategy—consistently yielded superior results. The VGG-16 architecture emerged as the most reliable classifier, an AUC-ROC of 98.59%, and a mean cross-validation accuracy of 98.92% ($\pm 0.08\%$). The VGG-19 variant performed comparably, yielding a cross-validation accuracy of 98.55% ($\pm 0.08\%$) and an AUC-ROC of 98.33%. Consequently, VGG-16 was selected for the final MCD module due to its high discriminative stability and F1-scores of 99.04% and 98.08% for negative and positive target classes, respec-

tively. Our refined VGG-16 ensemble significantly outperforms the 94% accuracy reported for pyramidal CNNs in Panama and the 90.2% accuracy achieved for African manatees using GoogLeNet [16,37]. This increased stability is primarily attributed to the bootstrap resampling strategy, which generated a balanced dataset of 100,000 spectrograms to mitigate the overfitting risks that limited prior AMCM iterations to 96% [25]. While transfer learning backbones pre-trained on high-diversity DBs provide superior throughput for distinguishing subtle vocalizations, a critical trade-off exists: VGG-based models incur higher inference latencies compared to the custom CNNs. This computational overhead may impact scalability in large-scale PAM deployments, suggesting that the custom CNN remains a viable alternative for real-time operations where execution speed is prioritized over marginal gains in detection quality.

Table 6. Comparative cross-validation execution time and performance metrics for supervised MCD architectures. The summary evaluates the different DL architectures compared in this study, across 10-folds using mean binary accuracy and binary cross-entropy loss, both with its standard deviation (**Std**). These results justify the selection of the VGG-16 architecture as the study's baseline [25], demonstrating superior error minimization and discriminative stability across several data partitions and combinations.

Model	Duration [s]	Accuracy (Std)	Loss (Std)
CNN	27.00×10^3	97.93% ($\pm 0.08\%$)	5.78% ($\pm 0.14\%$)
VGG-16	35.13×10^3	98.92% ($\pm 0.08\%$)	4.06% ($\pm 0.08\%$)
VGG-19	45.22×10^3	98.55% ($\pm 0.08\%$)	4.94% ($\pm 0.13\%$)

Subsequent deployment on unseen recordings from the Panamanian Caribbean confirmed the robust inference capabilities of the fine-tuned VGG-16 across diverse acoustic environments. However, elevated error rates in specific samples underscore a persistent challenge in PAM methodologies: the difficulty of isolating genuine biological signals from high-energy environmental or anthropogenic noise. Consequently, while metrics under controlled conditions are exceptional, they may overestimate generalization to independent field dataset, as a substantial portion of the training partition originates from captive environments that differ acoustically from wild conditions. This limitation, previously identified in the foundational AMCM framework, was further elucidated by the EDA visual suite. Manual inspection of the MCD inference revealed that several discrepancies stemmed from erroneous ground-truth annotations; in these instances, the model correctly identified positive vocalizations that had been overlooked during initial human labeling. Computational benchmarks revealed a distinct performance divergence between hardware architectures. Inference execution was markedly faster on the NVIDIA L40S-48GB, which leverages superior GPU memory capacity to accommodate large-scale models and extensive DBs. Conversely, the online WAV-FE pipeline demonstrated higher throughput on the legacy Tesla V100-PCIE-32GB. As detailed in Table 1, this behavior is justified by the V100 system's dual-CPU configuration and higher effective clock frequency (2.40 GHz), which better supports the serial, CPU-bound signal processing tasks inherent in the WAV-FE loop.

The unsupervised IMC module quantified vocalizations using the EM-KM algorithm, with the optimal cluster count ($k = 3$) determined via elbow method analysis of the inertia score. To maximize class separability, dimensionality reduction through PCA restricted the clustering space to the two most discriminative and positively correlated principal components. Despite a 25% counting error relative to the four individuals expected in the experimental set, the segregation of calls into three distinct clusters suggests the presence of individuals differentiated by unique acoustic signatures. This differentiation, supported

by acoustic KDE and spectral characterization, establishes a quantitative framework for demographic analysis, potentially facilitating life-stage classification (e.g., calves versus adults) based on F_0 and harmonic structures [17]. However, visual inspection of the bi-variate plot in Figure 10 suggests an additional latent cluster within the F_0 -BW feature space. Consequently, while the three identified clusters likely represent acoustically distinct individuals, these results should be interpreted as lower bounds of the actual population density. The EM-KM algorithm's imposition of spherical geometries may not fully capture the underlying acoustic variability. Alternative density-based or probabilistic approaches, such as HDBSCAN or GMMs, could reveal non-spherical structures and provide soft probabilistic assignments.

To contextualize the IMC module's performance, it is essential to compare it with studies where manatee identity was confirmed a priori. In the same geographic region of Bocas del Toro, prior research applied HDBSCAN clustering to vocalizations from 23 identified individuals, achieving an 83.75% correct assignment rate for scenarios involving ten individuals and at least thirty vocalizations per subject; however, accuracy inversely correlated with population size [19]. Similarly, HDBSCAN was utilized on zoo-recorded manatees to correctly identify population counts ranging from two to five individuals, though performance degraded when pooling data across facilities due to inconsistent noise conditions and varying vocalization counts [29]. Collectively, these findings indicate that while unsupervised clustering effectively resolves individual identity, it remains sensitive to population scale and acoustic environmental variability. This methodology yielded a silhouette coefficient of 79.20%, indicating superior cluster cohesion and separation relative to the 72% benchmark reported in previous iterations of this architecture [25]. This comparison suggests that the current framework provides a reliable approximation of individual presence, particularly when accounting for the inherent stochasticity of wild underwater environments.

Future Work

Future efforts prioritize the deployment of the proposed AMCM framework along the Costa Rican Caribbean coast. Utilizing an existing PAM-DB collected in field [25], the VGG-16 ensemble will perform call inference on unlabeled, long-term acoustic recordings from multiple protected areas. This application addresses the critical lack of ecological data and population status that currently hinders the development of conservation regulations for the Greater Caribbean manatee.

Analyzing these recordings will provide the scientific evidence—including occurrences, population density estimations and seasonal occurrence patterns—required to implement effective conservation strategies. The AMCM pipeline establishes a robust, scalable framework that accelerates reliable detection and population monitoring. Ultimately, this system facilitates high-accuracy ecological assessments, informing evidence-based conservation policies for environmental authorities in the region.

network (GAN) or variational auto-encoder (VAE) models, to increase the amount of labeled datasets. Despite the high-performance metrics attained in this study, the detection of FPs in field recordings confirms that environmental noise remains a persistent challenge for PAM-DBs. This generalization deficiency is primarily attributed to the reliance on replicated training data generated via bootstrap augmentation, which may produce harmonic configurations lacking the stochastic variability of natural acoustic signals. Consequently, future research must prioritize architectural robustness by refining training data quality through the integration of advanced generative AI frameworks for the synthesis of labeled samples without duplicates from authentic recordings. The implementation of generative adversarial network (GAN) or variational auto-encoder (VAE) architectures would

facilitate the creation of more realistic and variable synthetic spectral images, effectively bridging the gap between simulated benchmarks and complex field-recorded data.

network (GAN) or variational auto-encoder (VAE) models, to retrieve and integrate more realistic synthetic spectral images. Additionally, rigorous testing against local interference (e.g., boat noise) and the exploration of targeted denoising routines are essential to strengthen the model's performance against acoustic FPs, incorporating manual validation of specialist in marine bioacoustics, active learning step to verify proper detection of the MCD on large-scale PAM databases. Strengthening the model against acoustic interference requires rigorous evaluation of localized noise sources, such as vessel propulsion and snapping shrimp, alongside the exploration of targeted, adaptive denoising routines. Crucially, this development should establish an active learning framework, where detections manually validated by marine bioacoustic specialists are reincorporated into the training repository. This iterative refinement allows for the continuous fine-tuning of the model with expert input, ensuring its classification boundaries are dynamically adjusted to the evolving characteristics of real-world marine soundscapes and anthropocentric interference.

To further enhance the strategy for overfitting mitigation, the next crucial step involves the incorporation of automated hyperparameter tuning. Specifically, the Hyperband tuner will be utilized to efficiently explore the vast parameter space. This sophisticated approach automates the entire process of model configuration, effectively moving beyond subjective manual tuning or computationally expensive brute-force search methods, thereby ensuring an unbiased and optimal selection of model parameters for enhanced MCD performance.

Subsequent research must prioritize optimizing the clustering performance to enhance the silhouette coefficient and other key metrics. This optimization requires rigorous evaluation of alternative dimensionality reduction techniques, such as t-SNE or uniform manifold approximation and projection (UMAP), alongside other clustering algorithms, such as GMMs and HDBSCAN. These methods will be compared against the current approaches to achieve superior cluster cohesion and separation. As a further effort, this enhanced clustering capability is essential for investigating the integration of reliable acoustic individual identification within the IMC stage, a goal which remains an active and open topic in this field of research.

Author Contributions: Conceptualization, Quirós-Corella, F.Q.-C. and P.C.-P.; methodology, F.Q.-C., P.C.-P., and A.R.; software source code, F.Q.-C.; validation, F.Q.-C.; formal analysis, F.Q.-C.; investigation, F.Q.-C. and P.C.-P.; resources, P.C.-P., A.R. and B.B.; data curation, A.R. and B.B.; writing—original draft preparation, F.Q.-C.; writing—review and editing, A.R., B.B. and P.C.-P.; visualization, F.Q.-C.; supervision, F.Q.-C.; project administration, P.C.-P.; funding acquisition, P.C.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported and funded by the National Geographic (NatGeo) Society under project number NGS-84535T-21. Financial and computational support was provided by the AI4Earth program through grant number 69005a29-9390-4178-b6a2-4e4b68b470c6.

Institutional Review Board Statement: Data acquisition for Florida manatees under human care at ZooTampa was executed under formal authorization from the U.S. Fish and Wildlife Service (LOA #63658B). Similarly, research activities involving the manatee DTAG project was conducted in accordance with U.S. Fish and Wildlife Service permits #MA773494-8 and #MA773494-9. Furthermore, activities performed by the Florida Fish and Wildlife Conservation Commission (FWC) were exempt from licensing or registration requirements under the Animal Welfare Act. Field recordings utilizing passive acoustic recorders were reviewed and approved by the University of South Florida's Institutional Animal Care and Use Committee (IACUC) (W IS00007646), while research in Panama was conducted under the auspices of the Smithsonian Tropical Research Institute (STRI). Prior to the commencement of data acquisition, the project underwent comprehensive evaluation and received explicit authorization from the respective institutional research committees. These reviews ensured

that all recording procedures adhered to established animal welfare standards and institutional guidelines for non-invasive wildlife research, complying with the internal protocols and rigorous oversight mechanisms of all participating facilities.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study, comprising underwater acoustic recordings of marine mammals and the AI-based application source code, are available from the corresponding author upon reasonable request. Access to this data repository is restricted due to privacy, legal, and ethical considerations associated with working with vulnerable marine species such as the manatee. Specifically, open public access to these sensitive data carries the risk of misuse for unethical purposes, including manatee tracking or conducting studies that could cause harm or disturbance. This policy aligns with the principles of the 14th SDG by the UN, which commits to the protection of marine biodiversity for healthier oceans. Researchers interested in accessing the data are requested to contact the corresponding author to discuss the terms and conditions of data sharing. All requests will be evaluated on a case-by-case basis to ensure responsible and ethical utilization of the sensitive data repository.

Acknowledgments: We thank the Advanced Computing Laboratory at CeNAT for providing access to the Kabré high-performance computing infrastructure. We extend our gratitude to the ZooTampa project and the creators of the HaikuMarine system (David Mann and Austin Anderson) for their generous data sharing. Furthermore, we acknowledge the manatee DTAG project for sharing valuable data collected by the FWC staff and their partners. Finally, we thank Hector Guzmán and the STRI for providing access to experimental recordings collected in Bocas del Toro, Panama Caribbean, critically for validating manatee detection on real-world scenarios.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. May-Collado, L. Marine mammals. In *Marine Biodiversity of Costa Rica, Central America*; Springer: New York, NY, USA, 2009; pp. 479–495.
2. Keith Diagne, L. *Trichechus Senegalensis*. The IUCN Red List of Threatened Species 2015. Available online: https://www.researchgate.net/publication/350857049_THE_IUCN_RED_LIST_OF_THREATENED_SPECIES-_African_Manatee_Assessment_Errata_version (accessed on 15 February 2026).
3. Marsh, H. Dugong Dugon (Amended Version of 2015 Assessment). The IUCN Red List of Threatened Species 2019. 2019. Available online: <https://www.marinemammalhabitat.org/factsheets/northern-great-barrier-reef/> (accessed on 15 February 2026).
4. Freitas, K. Detecção de Zoonoses em Carnes de Caça Comercializadas na Região do Médio Rio Solimões—Coari-AM. *Instituto Nacional de Pesquisas da Amazônia—INPA* 2023. Available online: <https://www.gov.br/inpa/pt-br> (accessed on 15 February 2026).
5. Human activity Devastating Marine Species from Mammals to Corals—IUCN Red List, 2023. Available online: <https://iucn.org/press-release/202212/human-activity-devastating-marine-species-mammals-corals-iucn-red-list> (accessed on 15 February 2026).
6. Lin, M.; Turvey, S.T.; Han, C.; Huang, X.; Mazaris, A.D.; Liu, M.; Ma, H.; Yang, Z.; Tang, X.; Li, S. Functional extinction of dugongs in China. *R. Soc. Open Sci.* **2022**, *9*, 211994.
7. Marine Animals: Species Directory. Available online: <https://www.fisheries.noaa.gov/species-directory/marine-mammals> (accessed on 15 February 2026).
8. Kayanne, H.; Hara, T.; Arai, N.; Yamano, H.; Matsuda, H. Trajectory to local extinction of an isolated dugong population near Okinawa Island, Japan. *Sci. Rep.* **2022**, *12*, 6151.
9. Deutsch, C.; Self-Sullivan, C.; Mignucci-Giannoni, A. *Trichechus manatus*. The IUCN Red List of Threatened Species 2008: E. T22103A9356917. 2008. Available online: <https://manatipr.org/wp-content/uploads/2014/06/Deutsch08IUCN.pdf> (accessed on 15 February 2026).
10. Cubero-Pardo, P.; Castro-Azofeifa, C.; Corella, F.Q.; Ramírez, S.M.; Ramírez, E.V.; Sánchez, S.B.; Vargas-Bolaños, C. Antillean manatee (*Trichechus Manatus Manatus*) Occur. Grazing Spots Three Prot. Areas Costa Rica. *Lat. Am. J. Aquat. Mamm.* **2024**, *19*, 82–90.
11. Goal 14th: Life Below Water. 2024. Available online: <https://globalgoals.org/goals/14-life-below-water/> (accessed on 15 February 2026).

12. Ramos, E.A.; Maust-Mohl, M.; Collom, K.A.; Brady, B.; Gerstein, E.R.; Magnasco, M.O.; Reiss, D. The Antillean manatee produces broadband vocalizations with ultrasonic frequencies. *J. Acoust. Soc. Am.* **2020**, *147*, EL80–EL86.
13. Bittle, M.; Duncan, A. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. In *Proceedings of Acoustics*; Australian Acoustical Society: Victor Harbor, SA, Australia, 2013; Volume 2013.
14. Allen, A.N.; Harvey, M.; Harrell, L.; Jansen, A.; Merkens, K.P.; Wall, C.C.; Cattiau, J.; Oleson, E.M. A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Front. Mar. Sci.* **2021**, *8*, 607321.
15. Fleishman, E.; Cholewiak, D.; Gillespie, D.; Helble, T.; Klinck, H.; Nosal, E.M.; Roch, M.A. Ecological inferences about marine mammals from passive acoustic data. *Biol. Rev.* **2023**, *98*, 1633–1647.
16. Rycyk, A.M.; Berchem, C.; Marques, T.A. Estimating Florida manatee (*Trichechus Manatus Latirostris*) Abundance Using Passiv. Acoust. Methods. *JASA Express Lett.* **2022**, *2*, 051202.
17. Brady, B.; Ramos, E.A.; May-Collado, L.; Landrau-Giovannetti, N.; Lace, N.; Arreola, M.R.; Santos, G.M.; da Silva, V.M.F.; Sousa-Lima, R.S. Manatee calf call contour and acoustic structure varies by species and body size. *Sci. Rep.* **2022**, *12*, 19597.
18. Usman, A.M.; Ogundile, O.O.; Versfeld, D.J. Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access* **2020**, *8*, 105181–105206.
19. Merchan, F.; Echevers, G.; Poveda, H.; Sanchez-Galan, J.E.; Guzman, H.M. Detection and identification of manatee individual vocalizations in Panamanian wetlands using spectrogram clustering. *J. Acoust. Soc. Am.* **2019**, *146*, 1745–1757. [[CrossRef](#)]
20. Facheu, C.; Rycyk, A.M.; Kekeunou, S.; Keith-Diagne, L.W.; Ramos, E.A.; Kikuchi, M.; Takoukam Kamla, A. Acoustic methods improve the detection of the endangered African manatee. *Front. Mar. Sci.* **2023**, *9*, 1032464.
21. Erbs, F.; van der Schaar, M.; Marmontel, M.; Gaona, M.; Ramalho, E.; André, M. Amazonian manatee critical habitat revealed by artificial intelligence-based passive acoustic techniques. *Remote Sens. Ecol. Conserv.* **2024**, *11*, 172–186.
22. Sousa-Lima, R.S.; Paglia, A.P.; Da Fonseca, G.A. Signature information and individual recognition in the isolation calls of Amazonian manatees, *Trichechus Inunguis* (Mammalia: Sirenia). *Anim. Behav.* **2002**, *63*, 301–310.
23. Sousa-Lima, R.S.; Paglia, A.P.; da Fonseca, G.A.B. Gender, age, and identity in the isolation calls of Antillean manatees (*Trichechus Manatus Manatus*). *Aquat. Mamm.* **2008**, *34*, 109–122. [[CrossRef](#)]
24. Castro, J.M.; Rivera, M.; Camacho, A. Automatic manatee count using passive acoustics. In *Proceedings of Meetings on Acoustics*; Acoustical Society of America: Melville, NY, USA, 2015; Volume 23, p. 010001. [[CrossRef](#)]
25. Quirós-Corella, F.; Cubero-Pardo, P.; Rycyk, A.; Brady, B.; Castro-Azofeifa, C.; Mora-Ramírez, S.; Ureña-Madriral, J.P. An effective artificial intelligence pipeline for automatic manatee count using their tonal vocalizations. In *Proceedings of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Hernández-García, R.; Barrientos, R.J.; Velastin, S.A., Eds.; Springer: Cham, Switzerland, 2025; pp. 30–44.
26. Landrau-giovannetti, N.; Mignucci-giannoni, A.A.; Reidenberg, J.S. Acoustical and Anatomical Determination of Sound Production and Transmission in West Indian (*Trichechus Manatus*) Amaz. (*T. Inunguis*) Manatees. *Anat. Rec.* **2014**, *297*, 1896–1907.
27. Brady, B.; Moore, J.; Love, K. Behavior related vocalizations of the Florida manatee (*Trichechus Manatus Latirostris*). *Mar. Mammal Sci.* **2022**, *38*, 975–989.
28. O'Shea, T.J.; Poché, L.B. Aspects of underwater sound communication in Florida manatees (*Trichechus Manatus Latirostris*). *J. Mammal.* **2006**, *87*, 1061–1071. [[CrossRef](#)]
29. Schneider, S.; Von Fersen, L.; Dierkes, P.W. Acoustic estimation of the manatee population and classification of call categories using artificial intelligence. *Front. Conserv. Sci.* **2024**, *5*, 1405243.
30. Bianco, M.J.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M.A.; Gannot, S.; Deledalle, C.A. Machine Learning in acoustics: Theory and applications. *J. Acoust. Soc. Am.* **2019**, *146*, 3590–3628.
31. Mouy, X.; Leary, D.; Martin, B.; Laurinolli, M. A comparison of methods for the automatic classification of marine mammal vocalizations in the Arctic. In *Proceedings of the 2008 New Trends for Environmental Monitoring Using Passive Systems*, Hyeres, France, 14–17 October 2008; pp. 1–6.
32. Zhong, M.; Castellote, M.; Dodhia, R.; Lavista Ferres, J.; Keogh, M.; Brewer, A. Beluga whale acoustic signal classification using deep learning neural network models. *J. Acoust. Soc. Am.* **2020**, *147*, 1834–1841.
33. Liu, S.; Liu, M.; Wang, M.; Ma, T.; Qing, X. Classification of cetacean whistles based on convolutional neural network. In *Proceedings of the 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, Zhejiang, China, 18–20 October 2018; pp. 1–5.
34. Murphy, D.T.; Ioup, E.; Hoque, M.T.; Abdelguerfi, M. Residual learning for marine mammal classification. *IEEE Access* **2022**, *10*, 118409–118418.
35. Thomas, M.; Martin, B.; Kowarski, K.; Gaudet, B.; Matwin, S. Marine mammal species classification using convolutional neural networks and a novel acoustic representation. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Würzburg, Germany, 16–20 September 2019*; Springer: New York, NY, USA, 2019; pp. 290–305.
36. Lu, T.; Han, B.; Yu, F. Detection and classification of marine mammal sounds using AlexNet with transfer learning. *Ecol. Inform.* **2021**, *62*, 101277.

37. Merchan, F.; Guerra, A.; Poveda, H.; Guzmán, H.M.; Sanchez-Galan, J.E. Bioacoustic classification of Antillean manatee vocalization spectrograms using deep convolutional neural networks. *Appl. Sciences* **2020**, *10*, 3286.
38. Rycyk, A.; Bolaji, D.A.; Factheu, C.; Kamla Takoukam, A. Using transfer learning with a convolutional neural network to detect African manatee (*Trichechus Senegalensis*) Vocalizations. *JASA Express Lett.* **2022**, *2*, 121201.
39. Rycyk, A.; Cargille, V.; Bojali, D.; Factheu, C.; Ejimadu, U.; Berchem, C.; Takoukam Kamla, A. Bioacoustic dataset of African and Florida manatee vocalizations for Machine Learning applications, 2020-2022 ver 1. Environmental Data Initiative, 2025.
40. Rycyk, A.M.; Factheu, C.; Ramos, E.A.; Brady, B.A.; Kikuchi, M.; Nations, H.F.; Kapfer, K.; Hampton, C.M.; Garcia, E.R.; Takoukam Kamla, A. First characterization of vocalizations and passive acoustic monitoring of the vulnerable African manatee (*Trichechus Senegalensis*). *J. Acoust. Soc. Am.* **2021**, *150*, 3028–3037.
41. Knight, E.; Rhinehart, T.; de Zwaan, D.R.; Weldy, M.J.; Cartwright, M.; Hawley, S.H.; Larkin, J.L.; Lesmeister, D.; Bayne, E.; Kitzes, J. Individual identification in acoustic recordings. *Trends Ecol. Evol.* **2024**, *39*, 947–960.
42. Sainburg, T. Timsainb/Noisereduce: V1.0. 2019. Available online: <https://zenodo.org/records/3243139> (accessed on 15 February 2026).
43. Sainburg, T.; Thielk, M.; Gentner, T.Q. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput. Biol.* **2020**, *16*, e1008228.
44. Fitzgerald, D. Harmonic/percussive separation using median filtering. In Proceedings of the 13th International Conference on Digital Audio Effects (DAFX10), Graz, Austria, 6–10 September 2010.
45. Driedger, J.; Müller, M.; Disch, S. Extending harmonic-percussive separation of audio Signals. In Proceedings of the ISMIR, Taipei, Taiwan, 27–31 October 2014; pp. 611–616.
46. Zoubir, A.M.; Boashash, B. The bootstrap and its application in signal processing. *IEEE Signal Process. Mag.* **1998**, *15*, 56–76.
47. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.