International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00704-9

REGULAR PAPER

Automated cetacean detection in UAV imagery using AI models: a case study on Delphinid species

João Canelas[1,3] · Luana Clementino[2] · André Cid[3] · Joana Castro[3,4] · Inês Machado[2,5] · Susana Vieira[1]

Received: 30 June 2024 / Accepted: 13 December 2024
© The Author(s) 2024

Abstract

The identification and quantification of marine mammals is crucial for understanding their abundance, ecology and supporting their conservation efforts. Traditional methods for detecting cetaceans, however, are often labor-intensive and limited in their accuracy. To overcome these challenges, this work explores the use of convolutional neural networks (CNNs) as a tool for automating the detection of cetaceans through aerial images from unmanned aerial vehicles (UAVs). Additionally, the study proposes the use of Long-Short-Term-Memory (LSTM)-based models for video detection using a CNN-LSTM architecture. Models were trained on a selected dataset of dolphin examples acquired from 138 online videos with the aim of testing methods that hold potential for practical field monitoring. The approach was effectively validated on field data, suggesting that the method shows potential for further applications for operational settings. The results show that image-based detection methods are effective in the detection of dolphins from aerial UAV images, with the best-performing model, based on a ConvNext architecture, achieving high accuracy and f1-score values of 83.9% and 82.0%, respectively, within field observations conducted. However, video-based methods showed more difficulties in the detection task, as LSTM-based models struggled with generalization beyond their training environments, achieving a top accuracy of 68%. By reducing the labor required for cetacean detection, thus improving monitoring efficiency, this research provides a scalable approach that can support ongoing conservation efforts by enabling more robust data collection on cetacean populations.

Keywords Unmanned aerial vehicles · Convolutional neural networks · Long-short-term-memory · Machine learning · Marine mammals detection · Photo identification

1 Introduction

Cetaceans play a key role in maintaining ecosystem stability, acting as sentinel or indicator species that reflect the overall state of the ocean's health [1, 2]. Monitoring and safeguarding the diversity and abundance of cetaceans is imperative to support conservation efforts (e.g., through conventions and agreements) and achieve Good Environmental Status (GES) in European waters [3]. Achieving Good Environmental Sta-

Luana Clementino, André Cid, Joana Castro, Inês Machado and Susana Vieira have authors contributed equally to this work.

✉ João Canelas
joao.canelas@tecnico.ulisboa.pt

Luana Clementino
luana.clementino@wavec.org

André Cid
andre.cid@aimmportugal.org

Joana Castro
joana.castro@aimmportugal.org

Inês Machado
ines.machado@wavec.org

Susana Vieira
susana.vieira@tecnico.ulisboa.pt

[1] IDMEC, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

[2] WavEC Offshore Renewables, Edifício Diogo Cão, Doca de Alcântara Norte, 1350-352 Lisbon, Portugal

[3] AIMM - Associação para a Investigação do Meio Marinho, Rua Maestro Fred. Freitas N15-1, 1500-399 Lisbon, Portugal

[4] MARE - Marine and Environmental Sciences Centre/ARNET - Aquatic Research Network, Laboratório Marítimo da Guia, Faculdade de Ciências da Universidade de Lisboa, Av. Nossa Senhora do Cabo, 939, 2750-374 Cascais, Portugal

[5] MARE - Marine and Environmental Sciences Centre/ARNET - Aquatic Research Network, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal

Published online: 10 January 2025

tus (GES) in European waters is a key objective under the Marine Strategy Framework Directive (MSFD), which was adopted by the European Union to evaluate and maintain the health of the marine environment. GES is defined by eleven descriptors that assess various aspects of marine ecosystems, enabling a comprehensive evaluation of marine conditions and the pressures from human activities. This approach aligns with similar international conventions, such as the United Nations Sustainable Development Goal 14, which targets the conservation and sustainable use of oceans, seas, and marine resources, as well as the OSPAR Convention, which focuses on the protection of the North-East Atlantic marine environment. These frameworks collectively contribute to a more resilient and sustainably-managed global marine ecosystem. Monitoring and assessing the achievement of GES is particularly challenging given that cetaceans are highly mobile species, distributed over large areas, and moving across various marine habitats subject to diverse anthropogenic pressures. These pressures include incidental by-catch in fishing gear, bioaccumulation of pathogens and toxins, harmful algal blooms, collisions with ships, underwater noise and climate change [4–7]. More recently, the advancement of offshore renewable energy further intensified these challenges. Such projects often target large marine areas, commonly overlapping with cetacean habitats, thereby escalating the pressure between conservation needs and energy exploitation [8]. The vulnerability of these species and exploitation of their habitats underscores the importance of their conservation, thus, it is imperative to improve our current understanding of cetacean distribution patterns. However, such studies are excessively costly, posing a significant barrier to advancing conservation efforts. Traditional methods to study and monitor marine mammal populations involve visual surveys from a defined platform (e.g., aerial, ship-based, or land-based), acoustic surveys [10, 11], observation of Very High Resolution (VHR) satellite images [12, 13], and observation methods that allow a more thorough understanding (e.g., capture-recapture) [14]. Furthermore, emerging methodologies, such as remote sensing through photo detection and identification, present a promising tool for complementing such methods while reducing associated costs and risks [15–18].

Unmanned Aerial Vehicles (UAVs) are equipped with imaging sensors that can collect extremely high-resolution data, thus becoming an increasingly used tool for researchers to observe marine wildlife and study cetaceans. UAVs are a non-invasive method [19] that allows the detectability of animals in subsurface waters, thus increasing the time available for detection [20]. UAVs have an increasing number of applications, such as monitoring abundance and distribution [21], photo identification [22], behavioral studies [23], among others [20].

Nonetheless, this new technology still presents limitations and challenges, particularly associated with data management. The high volume of data generated requires efficient processing solutions, as manual inspection is often impractical and prone to human error [24, 25]. As machine learning and computer vision advance rapidly, automated computer vision models present a promising solution for automating the inspection process. Since the scientific developments by Krizhevsky et al. [26] proving the efficacy of deep learning algorithms in image recognition, Convolutional Neural Networks (CNNs) have become the model of choice for image detection and identification, achieving results on par with human performance in detection and identification tasks [27]. These models have been successfully applied to individual identification of whales [28–30] and dolphins, [31] with methods that can be adapted to other cetacean species [29]. There have also been applications for whale counting through satellite VHR images [32], where combining the usual counting procedure with an initial detection of whale presence has improved model accuracy and computational efficiency [32]. However, these are limited to species of larger size due to the spatial resolution of images and are difficult to develop due to the lack of open VHR image datasets.

Models developed for image detection, however, are limited to the events occurring in one single frame, potentially missing important information and context from the video sequences obtained with UAVs. That said, algorithms capable of handling video frames, such as Recurrent Neural Networks (RNNs), leverage the temporal continuity and contextual information provided by video sequences while reducing information missed, thus improving detection capability. Still, the development of such models represents a higher degree of complexity, and studies exploring their efficacy on marine mammal detection are relatively limited [33–35].

The main objective of this study is to develop machine learning models capable of automating the detection of cetaceans, specifically delphinids, using UAV data. The present work explores the implementation of well-documented CNN architectures directed to image detection, while also proposing the use of a Deep Fake detection algorithm based on Guera and Delp's "Deepfake video detection using recurrent neural networks" [36], applied to the detection of marine mammals in video sequences. This approach builds upon current methodologies, but also explores new avenues through the use of a Long-Short-Term-Memory (LSTM) network, a specific RNN model, seeking to harness the additional information provided through video analysis.

This study explores the synergies between deep learning and marine science, focusing on the potential to enhance environmental monitoring and impact assessment strategies in offshore environments, particularly for the conservation of dolphin populations. These findings provide a methodological basis for improving data quality, which can support future efforts aimed at advancing sustainable management and conservation efforts in marine ecosystems.

2 Data acquisition

In-situ collection of a sufficient volume of remote sensing data suitable for the development of an efficient identification model is challenging due to the high cost of equipment and logistical constraints associated with ocean surveys. Furthermore, datasets on species with broad distribution ranges, such as cetaceans, are scarce, and publicly available datasets tailored for aerial detection are largely non-existent. As a result, the data used to build the models were obtained by collecting scraped video files sourced from various online sources (e.g., YouTube, Pexels, Dailymotion, etc). Given the challenging nature of developing such datasets, data gathered for this study were limited to species of the family *Delphinidae*, as they are among the most accessible cetaceans in publicly available footage, and with the intent of gathering as much data as possible, no pre-selection criteria such as location or time period were applied during collection.

2.1 Training dataset

The videos obtained were recorded in diverse locations and under varying conditions, leading to significant variability in water characteristics such as hue, brightness, and foam, as well as differences among delphinid species. This diversity enhances the model's ability to generalize across different environmental settings. Furthermore, to achieve representative samples where no dolphins are present, the videos also include objects or subjects such as boats, boards, and swimmers which served as potential confounding elements for the model.

The resulting data consisted of 138 aerial videos of varying durations and settings. Some videos exclusively containing cetacean footage, others solely featuring water scenes, and some combining both elements. These 138 videos were then processed to create two distinct collections of data: one tailored for image classification and the other for video classification. For image classification, individual frames were extracted from the videos, providing static samples. For video classification, the original video segments were retained to capture dynamic features. While data for both methodologies were derived from the same set of videos, these were processed to suit the specific requirements of each classification type.

2.1.1 Image data

Image data were generated by deconstructing the original 138 raw videos into images by extracting frames at a specified rate of one frame every three seconds using the open software FFmpeg. This rate can be adjusted based on user needs and the source of extraction, as some videos may include more or less irrelevant data. In this study, images were categorized into two distinct classes, based on the presence or absence of cetaceans: "Cetacean" and "No Cetacean". Images were initially filtered to exclude the frames that were poor representatives of their class, such as cases where subjects were obstructed or not in the frame. Additional manual selection was also conducted on frames that were good representatives. The resulting set of data consisted of 2451 images, divided into its respective classes. The "No Cetacean" class included images where no cetaceans were present, as well as images with other surface or subsurface artifacts that could lead the model to incorrectly label them as containing a cetacean. Including these artifacts within the "No Cetacean" class helps to correct for potential false positives by exposing the model to non-cetacean images that may resemble cetaceans. The "Cetacean" class included images where at least one cetacean was present. The classification process resulted in 776 images (approximately 31.1%) representing the "No Cetacean" class, and 1720 images (about 68.9%) representing the "Cetacean" class.

The notable imbalance in the number of images per class is due to the limited variation in water surface patterns over time. Frames extracted within a few seconds of each other are often nearly identical, providing minimal additional value. On the other hand, a dataset heavily composed of cetacean images could bias model predictions, increasing the rate of false positives. To mitigate this, the number of images in the "No Cetacean" class was increased by artificially generating new sea images from existing ones. This was achieved by introducing random variations in brightness, hue, and saturation to all newly generated samples. Additionally, further transformations were applied with varying probabilities: sharpness enhancement (25%), random mirroring (25%), blurring (25%), random rotations (15%), and random cropping (30%).

The described set of transformations was applied a total of 944 times on randomly selected samples from the "No Cetacean" class, generating an additional 944 images. This augmentation was performed to equalize the number of samples with that of the "Cetacean" class. The resulting balanced data were composed of a total of 3420 images, with an equal distribution of 1720 (50%) images per class.

2.1.2 Video data

Video data were generated by deconstructing the same 138 raw videos into several smaller videos (clips) of eight seconds each and subsequently extracting a total of 64 frames from each of these smaller videos. The initial fragmentation process of the original videos was performed using the software Adobe Premiere Pro 2020, version 14.0 (Adobe Inc., San Jose, California). Firstly, intervals of eight seconds were manually selected to accurately represent each class. Simple transformations, such as mirroring, cropping, varying brightness, and hue, were applied to some of the samples to introduce variation. Each segment was then exported to create new video samples.

This video length was selected based on careful analysis of initial data acquired, balancing the goal of capturing comprehensive information on dolphin behavior and movement within a concise timeframe. This interval proved effective for segmenting original videos with frequent transitions and various added content such as overlays, logos, or artifacts that could otherwise cause unwanted model responses. A longer interval would have significantly reduced the number of usable samples, while a shorter window risked losing contextual details, as many segments showed minimal movement over brief durations. The eight-second length, therefore, provided an optimal compromise, enabling ample sample quantity while retaining sufficient information for model training.

After this segmentation, each clip is processed using Python's OpenCV library to extract frames at a rate of eight frames per second, resulting in a batch of images containing a total of 64 frames per clip. The choice of frame extraction rate allowed for capturing as much information on dolphin behavior variations over time, while minimizing the number of images.

The resulting data post-processing operations consisted of 1216 videos, of which 622 belong to the "Cetacean" class (approximately 51.2%), while the remaining 594 videos belong to the "No Cetacean" class (approximately 48.8%). This equates to 1216 batches of 64 images each, totaling 77824 images spanning both classes.

2.2 Test dataset

To monitor and understand model performance over the course of training, models are tested on data not involved in their learning process. This practice provides insights into expected performance and generalization by evaluating samples the model's parameters were not directly adjusted to, providing a general understanding of model progress and anticipated behavior within similar data samples.

In this study, the test data was derived from the original dataset outlined in Sect. 2.1, from which a small portion was retracted. This division creates two distinct subsets: training data, comprising 80% of the original dataset, used to teach the model to recognize class patterns, and test data, making up the remaining 20%, to verify the state of models. Empirical studies suggest optimal results when reserving 20–30% of data for testing while using 70–80% for training [37]. This separation was done randomly from all available samples while keeping a proportional number of samples from each class, resulting in 688 (20%) test and 2752 (80%) training samples for image classification, and 244 (20%) test and 972 (80%) training samples for video classification.

2.3 Validation dataset

The validation data for this study were provided by Associação para a Investigação do Meio Marinho (AIMM), which supported the research by supplying UAV data from previous expeditions. Data were acquired on the coastal region in south Portugal within the Faro district. Specifically, the study area is located approximately 12 km offshore from the coastline of Albufeira, extending into the Atlantic Ocean. This region is a significant habitat for various cetacean species, especially delphinids such as common dolphins (*Delphinus delphis*) and bottlenose dolphins (*Tursiops truncatus*) [38–40].

A total of seven campaigns conducted between 2022 and 2023 were analyzed. One campaign was included in the training data to better adapt to local environmental conditions and UAV settings, while the remaining six campaigns were used for evaluation. These surveys were conducted in the morning, between 10:30 and 12:00, under favorable sea conditions defined by a sea state of $\leqslant 3$ according to the Beaufort scale, swells < 1.5 m, good visibility (> 5 km), and no precipitation. Figure 1 offers a comprehensive view of the region under study, providing information on various expeditions, including dates, times, and the precise locations of dolphin sightings.

The UAV-based remote sensing data used in this study were collected using a Mavic 2 Pro multi-rotor UAV (DJI, Shenzhen, China). The UAV captured videos at a resolution of 3840×2160 pixels using a 1-inch CMOS RGB imaging sensor with a maximum resolution of 20-megapixel, coupled with a 3-axis gimbal and a 28 mm equivalent, f/2.8-f/11 lens, providing a field of view of approximately 77°.

The drone missions were conducted at different flight altitudes, depending on several factors. These factors included whether there were any dolphin sightings at the time and the size of the group of dolphins, with higher flights preferably used for greater sea coverage when no sightings were present, and lower flights for a more detailed view when a group was located. Figure 2 presents a box plot of the flight altitudes recorded by the UAV during the different expeditions. Of the six flights, three were conducted at a maximum altitude of

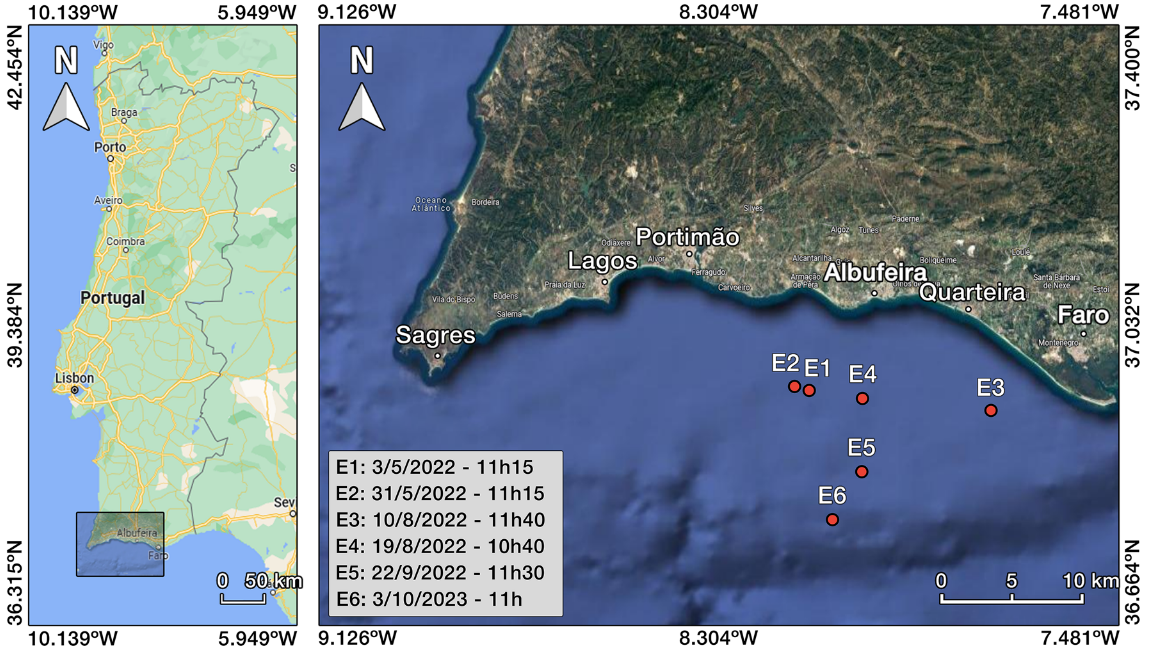

Fig. 1 Overview of the study area for acquiring data

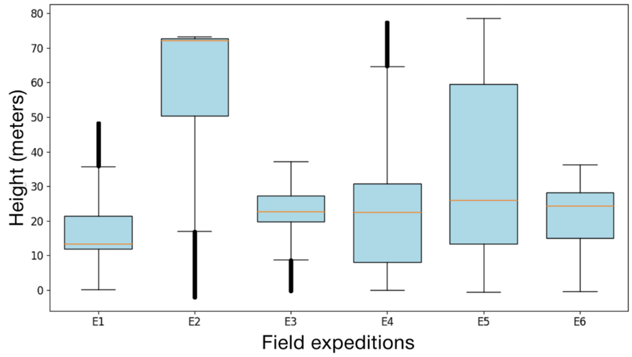

Fig. 2 Box plot: flight altitude distribution for each expedition

nearly 80 m, while the remaining three were flown below 50 m. In general, the UAV was observed to operate at an altitude of around 20 m, except for the second campaign, where flights predominantly occurred at higher altitudes.

The UAV-based imagery collected resulted in 35 min and 40 s of footage. Similar to previously acquired data, this footage was processed to create two datasets from the same source, this time for validation. The first, tailored for the validation of image-based models, was obtained by extracting and labeling frames from the original raw video data at a rate of one frame every five seconds, effectively processing the entire footage. Each sample was labeled as belonging to either the "Cetacean" or "No Cetacean" class based on inference from the original video imagery captured, allowing to discern the presence of dolphins on samples that would otherwise be challenging to identify correctly. The resulting data comprises a total of 428 image samples, with 247 (approximately 57.7%) classified as "Cetacean" and 181 (approximately 42.3%) as "No Cetacean". These samples capture a variety of dolphin positions, camera angles, and distances, providing sufficient variation to support a robust and comprehensive assessment. The second dataset, designed for the validation of video-based models, was obtained by manually dividing the original video data into smaller eight-second clips, extracting them and subsequently converting them into 64 images. The resulting video data consists of 232 videos, 120 of which were classified as belonging to the "Cetacean" class (approximately 51.7%), and 112 classified as belonging to the "No Cetacean" class (approximately 48.3%).

Table 1 provides a summary of the sample distribution across the training, testing, and evaluation datasets for both image and video data. Each dataset is divided into "Dolphin" and "Ocean" samples, corresponding to the "Cetacean" and "No Cetacean" classes, with the training and testing sets holding 80% and 20% of the original data, respectively. The evaluation dataset includes an additional set of samples that covers 100% of its allocated data, ensuring comprehensive assessment of the models. This division maintains a balanced class representation within each subset, with a near-equal distribution between "Dolphin" and "Ocean" samples across the datasets, facilitating robust training and performance evaluation.

While the evaluation dataset was enriched with cetacean images to ensure sufficient data for testing the model, it is acknowledged that in real-world applications, ocean-only images are likely to be far more prevalent than cetacean sightings. Consequently, this enrichment may slightly underestimate the rate of false positives under operational conditions. However, by maintaining a balanced dataset for evaluation, the accuracy metric becomes more representative of the model's true performance.

3 Implementation

The models and pipelines employed in this study were developed within the research platform Google Colab Pro, taking advantage of its cloud computing capabilities. The primary programming language used was Python version 3.9 with PyTorch's library as the foundation of this project's machine learning framework, which allowed for an easy implementation of state-of-the-art deep learning techniques.

3.1 Image-based models

In order to analyze distinct image identification models, the following CNN architectures were individually employed, as outlined in Table 2. The selected models have a track record in the field and known performance in image classification [41–45]. It is worth noting that while certain models may display superior performance on average, real-world outcomes can vary significantly based on the specific nature of the problems being addressed. Alongside the model names, specifications such as the number of parameters and the top accuracies achieved when these architectures were trained on ImageNet are also provided.

Initially, these models were set up according to their predefined architectures and initialized with random parameters, making them essentially empty frameworks incapable of making meaningful predictions. However, through transfer learning, parameters from models with identical architectures that have been trained on extensive datasets, such as ImageNet, can be transferred to these models. ImageNet, for example, comprises over a million samples and covers a wide range of classes, including animals like gray whales, dugongs, orcas, and sea lions. While it does not encompass the specific "dolphin" class, the features that distinguish these related classes can be invaluable for the identification of dolphins.

The models presented are structured in two sections, the first being the feature extractor, mainly consisting of the input layer and a series of hidden layers. It constitutes the majority of the network and is designed to identify specific features within the input data through its convolutional layers. These layers have been previously trained on the ImageNet dataset, therefore, they already possess the ability to effectively recognize a wide range of characteristics from a thousand different classes. Consequently, their parameters are "frozen" during further training to ensure that the extracted features remain consistent.

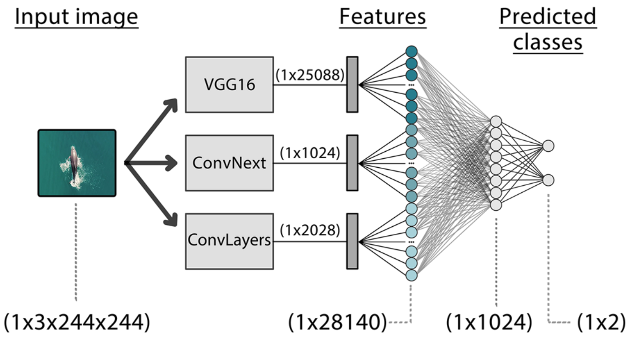

Fig. 3 Combined CNN model pipeline

The second section of the model comprises the classifier, typically a fully connected linear layer with a softmax activation function at the end of the network. This classifier interprets the features obtained from the hidden layers and assigns a class label to each data sample. The classifier, originally designed to classify 1000 classes, was adapted to identify only two classes. This was achieved by replacing the final linear layer, which initially had 1000 output nodes, with a new linear layer containing only two output nodes, corresponding to the two target classes.

In addition to the individual architectures, a combined model was developed to leverage multiple feature extractors concurrently. In this approach, the feature extraction layers from VGG16, ConvNext, and a straightforward set of convolutional layers were merged into a unified feature extractor. The simple convolutional model implemented along VGG16 and ConvNext consists of three convolutional layers and three max-pooling layers, complemented by certain Rectified Linear Unit (ReLU) layers in between as activation functions. The intent behind this design was to tailor the model to identify dolphin-specific features from the training dataset, instead of those that represent other subjects as in the case of transfer learning.

Ultimately, a shared classifier with two layers and 1024 nodes in its middle layer was employed to receive and effectively process the features extracted from all architectures, resulting in a collective prediction. Notably, this implementation was carried out in two instances: one that omitted the Convolutional Layers, CombinedModel (1), while the other used all three architectures as explained CombinedModel (2). Figure 3 provides a general overview of this combined model implementation and how data were shaped through it.

3.2 Video-based models

A video-based identification approach incorporating modern deepfake detection techniques was also adopted to leverage

Table 1 Dataset train-test split

Image dataset						Video dataset					
Train samples		Test samples		Eval samples		Train samples		Test samples		Eval samples	
2752 (80%)		688 (20%)		428 (100%)		972 (80%)		244 (20%)		232 (100%)	
Dolphin	Ocean	Dolphin	Ocean	Dolphin	Ocean	Dolphin	Ocean	Dolphin	Ocean	Dolphin	Ocean
1376	1367	194	194	247	181	497	475	125	119	120	112
(50%)	(50%)	(50%)	(50%)	(57.7%)	(42.3%)	(51.1%)	(48.9%)	(51.2%)	(48.8%)	(51.7%)	(48.3%)

Table 2 Common CNN architectures and respective performances on ImageNet dataset

Model	Top-1 Acc (%)	Top-5 Acc (%)	Parameters (M)
ResNet50	80.858	95.434	25.6
InceptionV3	77.294	93.450	27.2
VGG16_bn	73.360	91.516	138.4
ConvNext_Base	84.062	96.870	88.6
EfficientNet_V0	77.692	93.532	5.3

the unique temporal continuity feature of videos. Unlike images, videos are composed of a sequence containing numerous frames, where adjacent frames display a substantial correlation and temporal continuity. The method implemented, based on the work by Guera and Delp [36], involves using a CNN without a classifier to extract features from individual video frames and feed the resulting sequence of features into an LSTM to analyze patterns in their temporal evolution.

Two CNN architectures were used for this purpose: InceptionV4 and ConvNext. InceptionV3 architecture was replicated from the original work, while the ConvNext model architecture was selected based on results from the image-based classification methodology counterpart. In line with the previous approach, models were established based on their respective architectures, initialized with random parameters, and subsequently refined by transferring parameters from pre-trained models on ImageNet. Since the CNNs within this CNN-LSTM pipeline are used exclusively for feature extraction, their parameters were "frozen" to maintain the consistency of the extracted features. Simultaneously, the classifiers were removed, enabling direct passage of the features identified from the hidden layers to the LSTM. Different CNN architectures have specific input size requirements: InceptionV3 and ConvNext have input sizes of 299 × 299 and 224 × 224 pixels, respectively, and extract 2048 and 1024 features, respectively.

To accommodate the distinct feature structures obtained from each CNN architecture, two distinct LSTM architectures were developed. Each was designed to handle a specific input size for the transferred features, aligned with its corresponding CNN. Both models were created with two recurrent layers of 256 nodes each. This means that for each model, two LSTMs were stacked together to form a stacked LSTM, with the second taking in the outputs from the first to compute a new output at each time step. This setup enabled the LSTMs to iteratively produce 256 values, representing their hidden states, for every frame in the video sequence.

To conclude the CNN-LSTM pipeline, a classifier was introduced to process the output produced by the LSTM cell and make predictions. The classifier implemented features a linear layer with 16384 nodes on its input side. At each time step, the LSTM processes a frame from the input video, thus generating 256 values, which correspond to the 256 nodes in its hidden layers, representing the hidden state at that specific time step.

To maximize the amount of information used within the classifier, all the hidden states produced by the LSTM cell were aggregated. This aggregation results in a total of 16,384 nodes on the classifier's input side since all input videos are pre-processed to consist of 64 frames. Furthermore, its output layer consists of two nodes representing the two available classes and utilizes a softmax activation to convert the raw output values into probabilities.

Figure 4 provides a general overview of the pipeline created, its inner workings, and how the data were shaped through this system.

3.3 Pre-processing data

Effective pre-processing is essential for preparing the training and testing datasets. Key steps include organizing data into manageable batches and applying transformations compatible with pre-trained models, which optimize learning and enhance model performance.

To maximize computational efficiency and improve learning precision, all data samples within the training and testing datasets were grouped into batches of 64 samples each. This

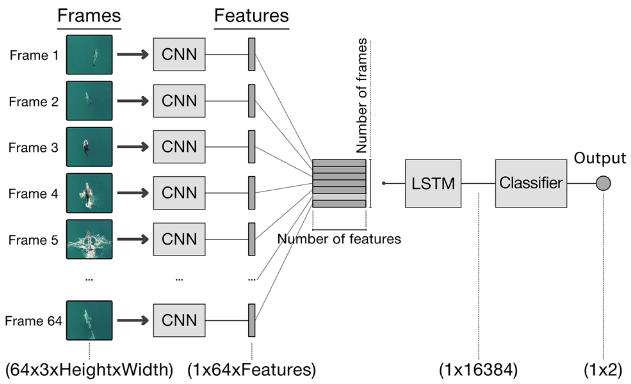

Fig. 4 CNN-LSTM pipeline structure

aggregation allows the simultaneous processing of multiple samples, providing a more stable gradient during backpropagation by combining losses from diverse samples within each batch. A batch size of 64, in particular, is a commonly employed choice that often works well for various deep learning tasks.

Further transformations were applied to leverage the patterns learned from pre-trained models. Given that weights from these were trained on data with a specific distribution, it is essential to normalize the input data accordingly. This normalization involves subtracting the mean and dividing by the standard deviation values of the dataset used to pre-train the models. For ImageNet-trained models, these values are (0.485, 0.456, 0.406) for the mean and (0.229, 0.224, 0.225) for the standard deviation across the three color channels. Following normalization, the input data are resized and cropped to match the dimensions required by each CNN architecture, with most models accepting (224 × 224) input, except InceptionV3, which requires (299 × 299).

3.4 Model training

The process of adapting a neural network to fit a specific problem involves iteratively assessing its performance on the training dataset, and adjusting its parameters each time to achieve predictions as close as possible to the desired values. To accomplish this, a Cross-Entropy Loss function, commonly utilized in multiclass problems such as this one was defined. This function serves as a guide to determine how the model's parameters should be updated. Subsequently, the loss function is utilized to compute a loss value, produced for each batch, which in turn is used to optimize the parameters of the model. This optimization is conducted through an Adam optimizer with an initial learning rate of 0.001, due to its adaptive learning rate and ease of use with fewer hyperparameters.

Additionally, a dropout layer with a dropout probability of 20% was added to the classifier at the end of each model during training, immediately before the final linear layer. This allows to randomly delete activations from the nodes carrying features before entering the classifier with a probability of 20% for each feature. This step proved to help the learning process of all nodes in the classifier and reduce data overfitting significantly by allowing nodes of undervalued importance to suffer larger adjustments.

Finally, models underwent training by iteratively processing the training dataset, one batch at a time, over several iterations, continuously assessing predictions made and adjusting their parameters accordingly. During this process, the dataset is completely processed multiple times, and models are stored for future use with their most up-to-date parameters and key performance metrics after each iteration.

4 Results and discussion

The predictive performance of the trained models was initially evaluated using the test dataset, offering an early indication of their effectiveness before validation on the evaluation dataset. This assessment includes metrics such as accuracy (Acc), precision (Prec), recall (Rec), f1-score (F1), and loss, providing a preliminary baseline of each model's generalization capacity. Table 3 summarizes the best-performing models, selected based on their f1-score, reflecting the balance between precision and recall.

Figure 5 shows the training curves for the top two models from each classification approach, highlighting the trends in loss and accuracy over epochs. These visualizations help clarify model stability and learning dynamics, setting the stage for a more detailed validation using the evaluation dataset in the subsequent section.

4.1 Model validation

To validate the effectiveness of the models studied and confirm the quality of their applicability, field observations were simulated using data collected during fieldwork conducted by AIMM. Subsequently, the performance of the pre-established models in training was assessed within the evaluation dataset detailed in Sect. 2.3. The quantitative results from this assessment are presented in Table 4.

The results show a clear fall in the overall performance of all models. This is to be expected since both training and testing data share the same origin source, and therefore, bear far more similarities. From the presented values it is possible to infer that the image-based identification achieved better performance than its video-based counterpart, establishing it as the superior methodology for this task within applied models.

Models based on the ConvNext architecture experienced a smaller drop in performance. In particular, the ConvNext

Table 3 Performance on test dataset

Model	Acc	Rec	Prec	F1	Loss
EfficientNet	0.834	0.875	0.809	0.841	0.006
ConvNext	0.969	0.965	0.974	0.969	0.001
ResNet50	0.898	0.916	0.885	0.900	0.004
InceptionV3	0.923	0.930	0.917	0.924	0.003
VGG16_bn	0.924	0.945	0.908	0.926	0.005
ConvolutionslLayers	0.850	0.863	0.841	0.852	0.018
CombinedModel (1)	0.956	0.965	0.949	0.957	0.004
CombinedModel (2)	0.955	0.945	0.964	0.954	0.006
CNN-LSTM (Incep.V3)	0.898	0.950	0.856	0.900	0.453
CNN-LSTM (ConvNext)	0.939	0.924	0.948	0.936	0.628

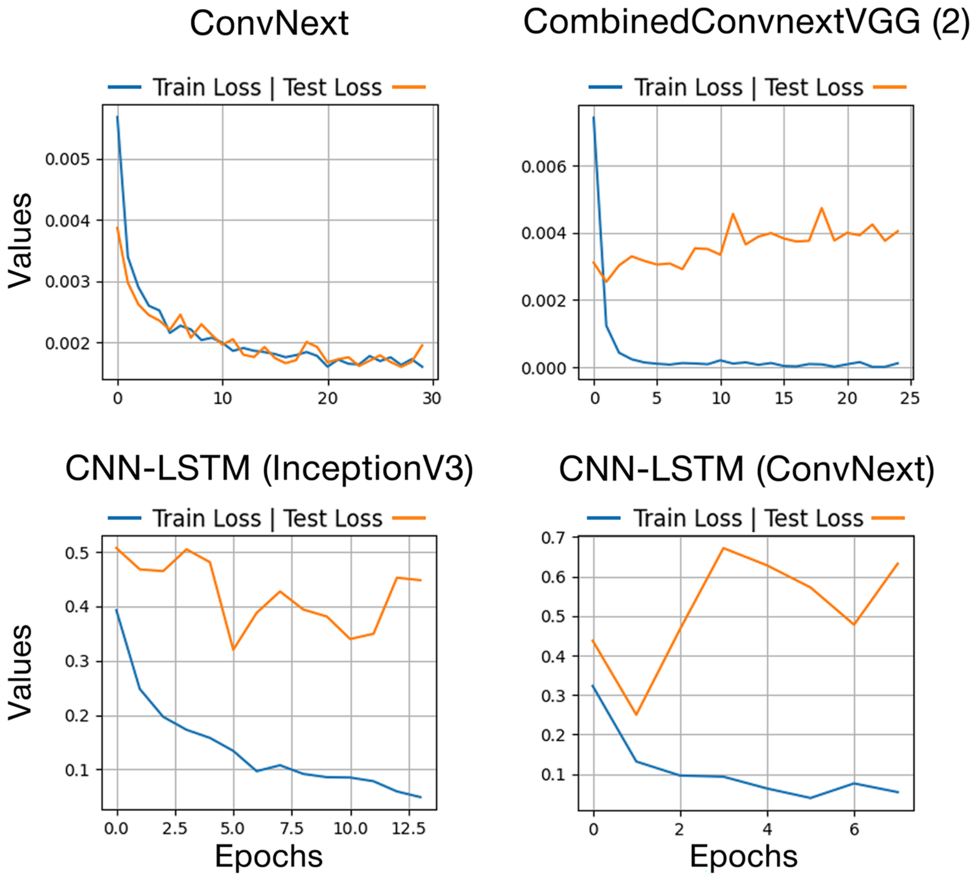

Fig. 5 Training curves from the most relevant models implemented

model stands out as the model of choice, achieving the highest accuracy and f1-score with values of 83.9% and 82.0%, respectively. This model is expected to produce overall good predictions, having achieved a better balance between recall and precision. Even though a decrease in recall was observed, the value of 86.7% still provides the model with a reduced number of false negatives, while the precision value of 77.7% leads to a reduced number of false positives compared to the remaining image-based models, meaning that predictions where the model finds the presence of dolphins are more trustworthy.

The notable performance of this model can be traced back to its training and testing curves, represented in Fig. 5, which, when compared to the curves of other models, display a higher degree of similarity, to the extent that they overlap. This suggests a great generalization ability by showing the model's capacity to obtain high accuracy values without overfitting training data.

Remaining ConvNext-based architectures have also demonstrated good performance. Notably, the performance of CombinedModel (2) improved compared to CombinedModel (1), achieving the highest recall value of 90.6%, which

Table 4 Performance on evaluation dataset

Model	Acc	Rec	Prec	F1	Loss
EfficientNet	0.678	0.906	0.575	0.704	0.801
ConvNext	0.839	0.867	0.777	0.820	0.355
ResNet50	0.759	0.873	0.664	0.754	0.503
InceptionV3	0.666	0.768	0.579	0.660	0.821
VGG16_bn	0.797	0.890	0.706	0.787	0.691
ConvolutionalLayers	0.661	0.707	0.582	0.638	1.676
CombinedModel (1)	0.799	0.892	0.706	0.788	0.680
CombinedModel (2)	0.811	0.906	0.719	0.802	1.203
CNN-LSTM (Incep.V3)	0.685	0.438	0.831	0.573	0.870
CNN-LSTM (ConvNext)	0.685	0.446	0.820	0.578	0.482

significantly reduces the number of false negatives. This improvement provides confidence that instances of dolphin presence are less likely to be missed, making the model more reliable for detecting such occurrences. This shift suggests an improvement in efficacy, potentially attributed to the convolutional layers specifically trained to identify dolphin features, which likely contributed to enhancing generalization capabilities in the model. Despite this, both models were still unable to surpass the performance of base ConvNext architecture, suggesting that the additional features from VGG16 and the extra trained convolutional layers did not provide a significant enhancement over the already proficient model.

In contrast, models like EfficientNet and InceptionV3 experienced significant drops in accuracy when tested on real-world settings, indicating a struggle with generalization, while models such as VGG16 and ResNet, though facing a moderate decrease in accuracy, still maintained relatively solid performance.

Additionally, the generalizability of models is shown to be likely connected to the number of parameters they contain, as larger models with more features appear to be less susceptible to overfitting in smaller datasets like the one used for training. In such a scenario, it is possible that the performance of the other architectures could be improved by utilizing their upscaled versions, which typically have more parameters and features, potentially enhancing their generalization capabilities. This is apparent when comparing Tables 2 and 5, where architectures with more parameters consistently outperformed the others. Even in the case of the VGG16 and ResNet50 models, where VGG16 which was initially expected to perform worse based on its ImageNet results, still outperformed ResNet50.

Regarding the video-based approach, CNN-LSTM models showed a significant drop in performance and displayed considerable bias toward the "No Cetacean" class, as can be observed by their recall values, indicating their struggles with generalization outside the environment they were trained. Consequently, although the 68.5% accuracy achieved might seem reasonable at first, this value does not provide an accurate representation of this methodology's efficacy in the study, as the performance of these models appears to be inversely proportional to their training. Figure 6 demonstrates the evolution in performance of the CNN-LSTM (InceptionV3) in the evaluation dataset as it trains further in the training set. From this, it is possible to observe that the model achieved a far better performance on its first iteration, and as it trains further there is a consistent decrease in accuracy, recall, and an increase in loss. These results show that the model is overfitting just after its first passage on the training dataset. This is also suggested by its training curves, presented in Fig. 5, which shows a clear and increasingly pronounced difference in training and testing loss. These models appear to be highly sensitive to variations in the evaluation dataset and may require further investigation and fine-tuning.

The causes of such overfitting are uncertain, nonetheless, a few reasons can be pointed out. As LSTMs are more complex than regular RNNs and tend to require a larger amount of training data to learn effectively, it is possible that a dataset containing only 972 training samples (refer to Table 1) may not suffice. Another possibility could be related to the structure of the model, which due to its propensity to memorize long-term dependencies that might not generalize well and overfit, may simply require further tuning. Finally, the use of individual, unbatched videos for training was done because each video already constitutes a batch of 64 frames, however, this could also be a contributing factor to overfitting on each sample.

4.2 Model applicability

This section presents practical examples of model performance when applied to evaluation data, and inspecting individual samples. The model of choice for this evaluation is ConvNext, having achieved superior performance compared to the remaining implementations. These examples reveal the

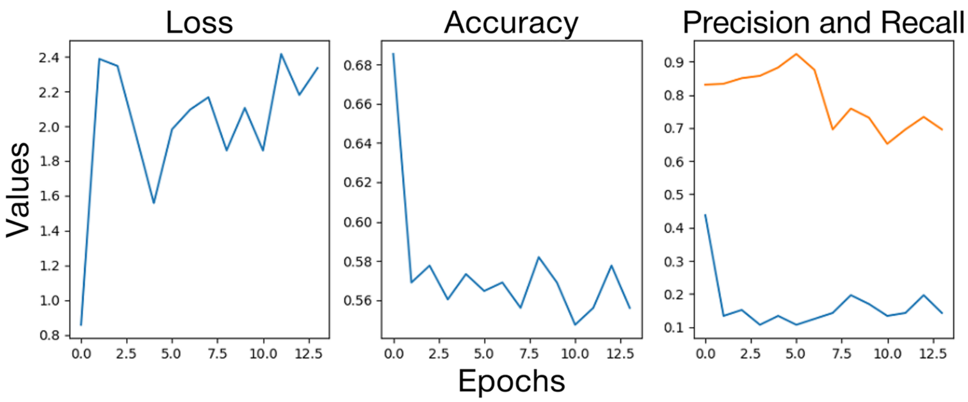

Fig. 6 CNN-LSTM performance on the validation set as trains further

Table 5 Distribution of predictions across the evaluation dataset

Samples	Beaufort	Dolphin	Ocean	Accuracy
E1	2	6/6	11/18	1.00 | 0.61
E2	3	4/15	45/46	0.27 | 0.98
E3	3	86/96	17/26	0.90 | 0.65
E4	1	0/0	11/11	– | 1.00
E5	0	0/0	37/38	– | 0.97
E6	1	106/130	36/42	0.82 | 0.86
Total	–	202/247	157/181	0.82 | 0.87

model's behavior in real-world scenarios, showcasing both its strengths and weaknesses, along with contextual factors affecting its performance.

Table 5 provides a comprehensive overview of the sample distribution, and the predictions generated by the model across various evaluation samples collected from different field expeditions. Additionally, the sea conditions, as described by the Beaufort scale, are included to offer insights into how environmental factors may have influenced the model's performance. The overall distribution of correct predictions, as well as respective accuracy levels, are presented at the bottom of the table for both classes, with the model having only missed 45 Dolphin and 24 Ocean samples, representing a total of 69 failed predictions out of 428 samples.

Ocean predictions remained consistent across the varying conditions of each expedition, with the largest performance drops occurring in the first and third expeditions. Dolphin predictions, however, struggled particularly during the second expedition, where the model missed 11 out of 15 dolphin samples, accounting for most of the false negatives proportionally. This trend confirms a correlation between model performance and sea conditions, as samples with a Beaufort value of 2 or above presented greater challenges for detection.

Another highly important factor when discussing aerial imagery is the resolution of data, as the amount of information within each pixel will vary significantly depending on factors such as camera angle and flight altitude. Figure 7 showcases the effect of this resolution on evaluated samples, displaying predictions and their confidence level based on their estimated Ground Sampling Distance (GSD), a common metric to evaluate the physical resolution of a pixel. Dolphin samples are represented by an "X" marking, with green indicating a correct prediction and red indicating an incorrect prediction. Similarly, Ocean samples are represented by a circle, with green and red denoting correct and incorrect predictions, respectively.

From this can be observed that the model displays a greater tendency toward false negatives for GSD values above 11 mm/pixel, with the model showing an accuracy and recall value of 86.4% and 82.6%, respectively, for GSD below this threshold, and only accuracy and recall values of 79.6% and 72.7%, respectively, for higher GSD values. This corresponds to a flight altitude of approximately 80 m with the camera pointing directly downward for the equipment used. However, this setup was not commonly employed in most of the data collection, as it generated excessive glare during the time of day when the surveys took place. Additionally, the model failed to correctly identify any dolphin samples for GSD values exceeding 20 mm/pixel, missing all 8 dolphin samples under these conditions.

Additional challenges, typically related to sea state, in the identification process, include the presence of significant reflections on the water surface and lack of brightness. Significant reflection may induce models in error, biasing predictions toward the "Cetacean" class, as observed by results from the first and third excursions which had a considerable amount of glare. Moreover, the overall results presented may overestimate the true predictive accuracy of the model due to the nature of the validation data, as evidenced by the lower observed precision value.

Human observers, however, face notable challenges in detecting animals during surveys. Factors such as fatigue during extended observation periods [46], restricted field of view, and difficulties in detecting submerged animals [47] significantly hinder detection rates. Observers often under-

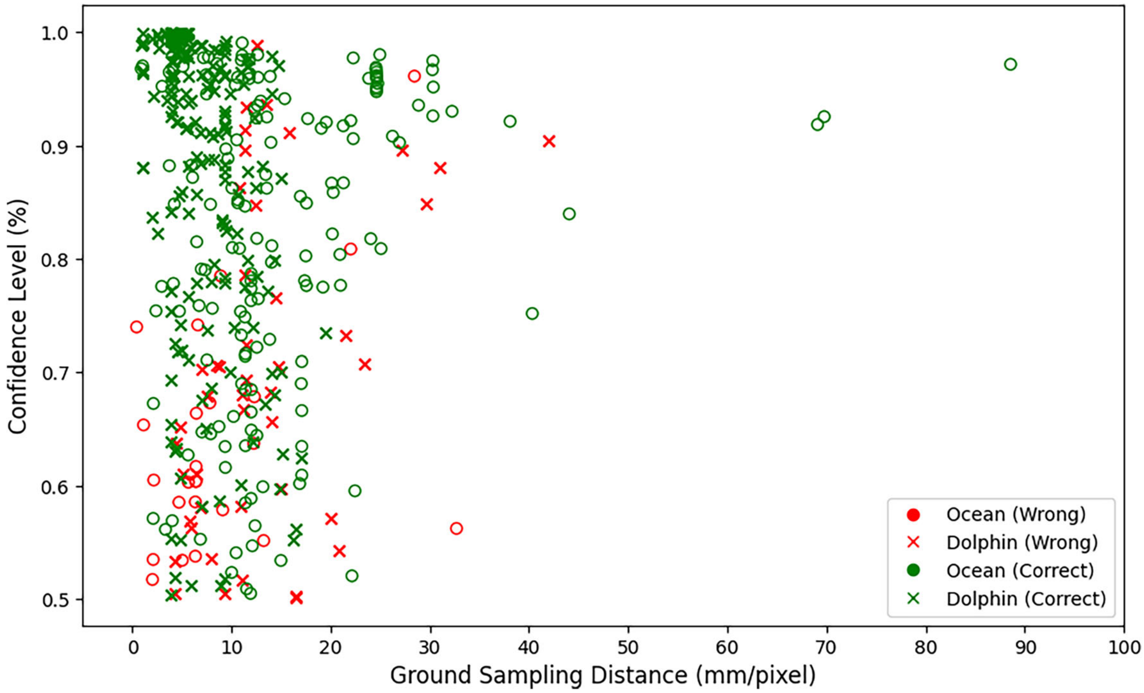

Fig. 7 Impact of ground sampling distance (GSD) on model predictions

estimate group sizes due to asynchronous diving behaviors [48], and perception bias can lead to misidentified sightings, as noted in studies reporting up to 5% detection errors [46].

In contrast, UAV methods, such as the one proposed here, demonstrate superior performance. According to Fettermann et al. (2022), UAVs identified larger group sizes in 71.4% of cases and detected 26.4% more individuals on average compared to human efforts [47]. Similarly, the model presented exhibits strong performance, offering significant reductions in manpower requirements while improving consistency and accuracy in wildlife monitoring.

5 Further developments

Future developments in image-based methods may also investigate different methodologies. One such approach involves exploring upscaled versions of the models used, as scalability proved to be a notable factor in model performance. Future investigations could try to improve upon current performance measurements and assess the potential trade-off between computing speed and model performance for different use cases such as real-time detection.

Further research aims to develop and improve object detection models, such as You Only Look Once (YOLO) combined with a tracking head such as Simple Online and Real-time Tracking (SORT), to enable tracking and counting of individuals. These developments could ultimately fit into a comprehensive pipeline capable of determining crucial information for marine researchers such as group sizes, distinguishing between juveniles and adults, and, at a later stage, aiding in species identification.

Finally, one of the primary limitations in this study is the size of the dataset, a potential reason for the suboptimal performance of some models, particularly video-based models. The datasets used in similar studies are often considerably larger, allowing for more thorough training. This challenge was addressed by enriching the dataset, to achieve a more balanced number of samples from the two classes, as well as increasing the dataset. It is worth noting that there is a potential risk of increasing the rate of false positives in enriched datasets, however, the models from the present study performed well and were able to achieve high accuracy in images without dolphin presence. Therefore, future efforts should focus on expanding the available data resources to enhance model performance and robustness. This will enable more comprehensive training and better generalization, ultimately improving the effectiveness of the models. Additionally, there were limitations associated with obtaining the dataset from online resources, namely the lack of detailed information on flight altitude and camera positioning. These factors directly impact image resolution and, consequently, model performance. In this study, these limitations were partially addressed through pre-processing techniques and model adjustments aimed at increasing robustness under varied conditions. However, future model developments will involve collecting a dataset with new imagery that documents altitude, camera positioning, and other relevant parameters.

This approach will ensure a robust and representative training dataset, allowing the model's performance to be effectively assessed and improved, reducing potential biases related to altitude and operational variability.

6 Conclusion

This research successfully investigated the use of machine learning models in the task of automatically detecting dolphins through aerial imagery.

The evaluation of the models demonstrated several key findings. Notably, image-based models developed from a training dataset created from data collected from a wide range of internet sources proved to be capable of yielding good results in classifying the presence of dolphins. Furthermore, models based on the ConvNext architecture exhibited a more robust performance compared to other architectures on all datasets and consistently demonstrated superior performance across all metrics. Particularly, the combined model CombinedModel (2), boasting additional convolutional layers trained to specifically identify dolphin features, outperformed its counterpart, suggesting an improvement in generalizability provided by these layers. On the other hand, suboptimal performance was observed from CNN-LSTM-based models in terms of generalization beyond their training environments. This sensitivity to variations in the evaluation dataset highlights the need for further investigation and fine-tuning. Ultimately, the ConvNext model emerged as a standout performer with 83.9% accuracy, 86.7% recall, and 77.7 precision.

A critical observation was on Filming conditions and sea states, which proved crucial to model performance. Results indicate that resolution values below 11 mm/pixel consistently improve performance within the given data. Additionally, surveys conducted under a Beaufort state of 2 yielded significantly better results. Factors such as glare on the water surface and the presence of boats were also observed to bias model predictions.

Overall, the results obtained provide practical insights that can guide the development of more robust and versatile models for the conservation of marine life, offering a promising direction for the future of aerial monitoring in marine environments.

Acknowledgements The authors acknowledge Fundação para a Ciência e a Tecnologia (FCT) for its financial support via the projects LAETA Base Funding (DOI: 10.54499/UIDB/50022/2020) and LAETA Programatic Funding (DOI: 10.54499/UIDP/50022/2020). The authors also acknowledge the EU-SCORES project, that received funding from the European Union's Horizon 2020 research innovation programme under the Grant Agreement n° 101036457. This work was conducted within the Department of Mechanical Engineering (DMEC) at Instituto Superior Técnico.

Author Contributions I. Machado conceptualized the study. J. Canelas and S. Vieira defined the methodology. J. Canelas performed data curation by scraping online video data for training and analyzed all data. A. Cid collected the validation data curated by J. Castro. J. Canelas conducted the formal analysis and investigation, and visualized the results. J. Canelas and L. Clementino drafted the manuscript, which was reviewed and edited by L. Clementino, I. Machado, J. Castro and S. Vieira. S. Vieira and I. Machado supervised the work, with J. Castro providing validation data as a resource. All authors approved the final manuscript for publication.

Funding This work was funded by Fundação para a Ciência e a Tecnologia (FCT) via the projects LAETA Base Funding (DOI: 10.54499/UIDB/50022/2020) and LAETA Programatic Funding (DOI: 10.54499/UIDP/50022/2020). This work was supported by the EU-SCORES project, that received funding from the European Union's Horizon 2020 research innovation programme under the Grant Agreement n° 101036457; this study was funded by Associação para a Investigação do Meio Marinho (AIMM)

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval and consent to participate All components of the data collection were approved by the Portuguese Conservation Institute. This work was conducted under the scientific permit from the Portuguese Conservation Institute (ICNF: AOC/34/2020), and all animal protocols adhered to their standards.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Bowen, W.D.: Role of marine mammals in aquatic ecosystems. Mar. Ecol. Prog. Ser. **158**, 267–274 (1997)
2. Sergio, F., Caro, T., Brown, D., Clucas, B., Hunter, J., Ketchum, J., McHugh, K., Hiraldo, F.: Top predators as conservation tools: ecological rationale, assumptions, and efficacy. Annu. Rev. Ecol. Evol. Syst. **39**, 1–19 (2008)
3. Directive, S.F.: Directive 2008/56/ec of the European parliament and of the council. Journal. Council Decision of (2008)
4. Parsons, E., et al.: An introduction to marine mammal biology and conservation (2013)
5. Albouy, C., Delattre, V., Donati, G., Frölicher, T.L., Albouy-Boyer, S., Rufino, M., Pellissier, L., Mouillot, D., Leprieur, F.: Global

vulnerability of marine mammals to global warming. Sci. Rep. **10**(1), 548 (2020)

6. Avila, I.C., Kaschner, K., Dormann, C.F.: Current global risks to marine mammals: taking stock of the threats. Biol. Cons. **221**, 44–58 (2018)
7. Duarte, C.M., Chapuis, L., Collin, S.P., Costa, D.P., Devassy, R.P., Eguiluz, V.M., Erbe, C., Gordon, T.A., Halpern, B.S., Harding, H.R., et al.: The soundscape of the Anthropocene ocean. Science **371**(6529), 4658 (2021)
8. Bald, J.: Reviewing the ecological impacts of offshore wind farms. npj Ocean Sustain. **1**, 1–8 (2022)
9. Iglesias, G., Tercero, J.A., Simas, T., Machado, I., Cruz, E.: Environmental effects. In: Greaves, D., Iglesias, G (eds) Wave and tidal energy (2018). https://doi.org/10.1002/9781119014492.ch9
10. Marques, T.A., Munger, L., Thomas, L., Wiggins, S., Hildebrand, J.A.: Estimating north pacific right whale eubalaena japonica density using passive acoustic cue counting. Endangered Species Res. **13**(3), 163–172 (2011)
11. Mellinger, D.K., Stafford, K.M., Moore, S.E., Dziak, R.P., Matsumoto, H.: An overview of fixed passive acoustic observation methods for cetaceans. Oceanography **20**(4), 36–45 (2007)
12. Kennedy, A.S., Zerbini, A.N., Vásquez, O., Gandilhon, N., Clapham, P.J., Adam, O.: Local and migratory movements of humpback whales (megaptera novaeangliae) satellite-tracked in the north Atlantic ocean. Can. J. Zool. **92**(1), 9–18 (2014)
13. Wade, P., Heide-Jørgensen, M.P., Shelden, K., Barlow, J., Carretta, J., Durban, J., LeDuc, R., Munger, L., Rankin, S., Sauter, A., et al.: Acoustic detection and satellite-tracking leads to discovery of rare concentration of endangered north pacific right whales. Biol. Let. **2**(3), 417–419 (2006)
14. Bertulli, C.G., Guéry, L., McGinty, N., Suzuki, A., Brannan, N., Marques, T., Rasmussen, M.H., Gimenez, O.: Capture-recapture abundance and survival estimates of three cetacean species in Icelandic coastal waters using trained scientist-volunteers. J. Sea Res. **131**, 22–31 (2018)
15. Hodgson, A., Kelly, N., Peel, D.: Unmanned aerial vehicles (UAVS) for surveying marine fauna: a dugong case study. PLoS ONE **8**(11), 79556 (2013)
16. Hodgson, J.C., Baylis, S.M., Mott, R., Herrod, A., Clarke, R.H.: Precision wildlife monitoring using unmanned aerial vehicles. Sci. Rep. **6**(1), 22574 (2016)
17. Rodofili, E.N., Lecours, V., LaRue, M.: Remote sensing techniques for automated marine mammals detection: a review of methods and current challenges. PeerJ **10**, 13540 (2022)
18. Winkler, C., Panigada, S., Murphy, S., Ritter, F.: Global numbers of ship strikes: an assessment of collisions between vessels and cetaceans using available data in the IWC ship strike database. IWC B **68**, 66 (2020)
19. Castro, J., Borges, F.O., Cid, A., Laborde, M.I., Rosa, R., Pearson, H.C.: Assessing the behavioural responses of small cetaceans to unmanned aerial vehicles. Remote Sens. **13**(1), 156 (2021)
20. Álvarez-González, M., Suarez-Bregua, P., Pierce, G.J., Saavedra, C.: Unmanned aerial vehicles (UAVS) in marine mammal research: a review of current applications and challenges. Drones **7**(11), 667 (2023)
21. Cleguer, C., Kelly, N., Tyne, J., Wieser, M., Peel, D., Hodgson, A.: A novel method for using small unoccupied aerial vehicles to survey wildlife species and model their density distribution. Front. Mar. Sci. **8**, 640338 (2021)
22. Ryan, K.P., Ferguson, S.H., Koski, W.R., Young, B.G., Roth, J.D., Watt, C.A.: Use of drones for the creation and development of a photographic identification catalogue for an endangered whale population. Arctic Sci. **8**(4), 1191–1201 (2022)
23. Torres, L.G., Nieukirk, S.L., Lemos, L., Chandler, T.E.: Drone up! quantifying whale behavior from a new perspective improves observational capacity. Front. Mar. Sci. **5**, 319 (2018)
24. Chabot, D., Stapleton, S., Francis, C.M.: Using web images to train a deep neural network to detect sparsely distributed wildlife in large volumes of remotely sensed imagery: A case study of polar bears on sea ice. Eco. Inform. **68**, 101547 (2022)
25. Maksimenko, V.A., Hramov, A.E., Frolov, N.S., Lüttjohann, A., Nedaivozov, V.O., Grubov, V.V., Runnova, A.E., Makarov, V.V., Kurths, J., Pisarchik, A.N.: Increasing human performance by sharing cognitive load using brain-to-brain interface. Front. Neurosci. **12**, 949 (2018)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28**, 66 (2015)
28. Blount, D., Holmberg, J., Parham, J., Gero, S., Gordon, J., Levenson, J.J.: Comparison of three individual identification algorithms for sperm whales (physeter macrocephalus) after automated detection. bioRxiv, 2021–12 (2021)
29. Bogucki, R., Cygan, M., Khan, C.B., Klimek, M., Milczek, J.K., Mucha, M.: Applying deep learning to right whale photo identification. Conserv. Biol. **33**(3), 676–684 (2019)
30. Hiby, L., Lovell, P., et al.: A note on an automated system for matching the callosity patterns on aerial photographs of southern right whales. J. Cetacean Res. Manage. **66**, 291–295 (2020)
31. Patton, P.T., Cheeseman, T., Abe, K., Yamaguchi, T., Reade, W., Southerland, K., Howard, A., Oleson, E.M., Allen, J.B., Ashe, E., et al.: A deep learning approach to photo-identification demonstrates high performance on two dozen cetacean species. Methods Ecol. Evol. **14**(10), 2611–2625 (2023)
32. Guirado, E., Tabik, S., Rivas, M.L., Alcaraz-Segura, D., Herrera, F.: Whale counting in satellite and aerial images with deep learning. Sci. Rep. **9**(1), 1–12 (2019)
33. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 803–818 (2018)
34. Sigurdsson, G.A., Russakovsky, O., Gupta, A.: What actions are needed for understanding human actions in videos? In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2137–2146 (2017)
35. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
36. Güera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2018). IEEE
37. Gholamy, A., Kreinovich, V., Kosheleva, O.: Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. Int. J. Intell. Technol. Appl. Stat. **11**(2), 105–111 (2018)
38. Morais, P., Afonso, L., Dias, E.: Harnessing the power of social media to obtain biodiversity data about cetaceans in a poorly monitored area. Front. Mar. Sci. **8**, 765228 (2021)
39. Castro, J.M.C.: Characterization of cetaceans in the south coast of portugal between lagos and cape são vicente. Master's thesis, Universidade de Lisboa (Portugal) (2010)
40. Castro, J., Faustino, C., Cid, A., Quirin, A., Matos, F.L., Rosa, R., Pearson, H.C.: Common dolphin (*Delphinus delphis*) fission-fusion dynamics in the south coast of Portugal. Behav. Ecol. Sociobiol. **76**(9), 128 (2022)
41. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
42. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
43. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
44. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
46. Oliveira-Rodrigues, C., Correia, A.M., Valente, R., Gil, Á., Gandra, M., Liberal, M., Rosso, M., Pierce, G., Sousa-Pinto, I.: Assessing data bias in visual surveys from a cetacean monitoring programme. Sci. Data **9**(1), 682 (2022)
47. Fettermann, T., Fiori, L., Gillman, L., Stockin, K.A., Bollard, B.: Drone surveys are more accurate than boat-based surveys of bottlenose dolphins (tursiops truncatus). Drones **6**(4), 82 (2022)
48. Brown, A.M., Allen, S.J., Kelly, N., Hodgson, A.J.: Using unoccupied aerial vehicles to estimate availability and group size error for aerial surveys of coastal dolphins. Remote Sens. Ecol. Conserv. **9**(3), 340–353 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.