Quo Vadimus

Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook[†]

Morten Goodwin[1], Kim Tallaksen Halvorsen [2], Lei Jiao[1], Kristian Muri Knausgård [3],
Angela Helen Martin [4], Marta Moyano [4], Rebekah A. Oomen [1,4,5,*], Jeppe Have Rasmussen[1,4],
Tonje Knutsen Sørdalen [4], and Susanna Huneide Thorbjørnsen[2,4]

[1]Centre for Artificial Intelligence Research, University of Agder, 4604 Kristiansand, Norway
[2]Institute of Marine Research, Nye Flødevigveien 20, Flødevigen, 4817 His, Norway
[3]Top Research Centre Mechatronics, University of Agder, 4879 Grimstad, Norway
[4]Center for Coastal Research, University of Agder, 4604 Kristiansand, Norway
[5]Center for Ecological and Evolutionary Synthesis, University of Oslo, 0371 Oslo, Norway

[*]Corresponding author: tel: +47 (47 71 44 17); e-mail: rebekahoomen@gmail.com
[†]All authors have an equal contribution.

The deep learning (DL) revolution is touching all scientific disciplines and corners of our lives as a means of harnessing the power of big data. Marine ecology is no exception. New methods provide analysis of data from sensors, cameras, and acoustic recorders, even in real time, in ways that are reproducible and rapid. Off-the-shelf algorithms find, count, and classify species from digital images or video and detect cryptic patterns in noisy data. These endeavours require collaboration across ecological and data science disciplines, which can be challenging to initiate. To promote the use of DL towards ecosystem-based management of the sea, this paper aims to bridge the gap between marine ecologists and computer scientists. We provide insight into popular DL approaches for ecological data analysis, focusing on supervised learning techniques with deep neural networks, and illustrate challenges and opportunities through established and emerging applications of DL to marine ecology. We present case studies on plankton, fish, marine mammals, pollution, and nutrient cycling that involve object detection, classification, tracking, and segmentation of visualized data. We conclude with a broad outlook of the field's opportunities and challenges, including potential technological advances and issues with managing complex data sets.

Keywords: artificial intelligence, ecosystem-based management, machine learning, marine bioacoustics, marine monitoring.

Introduction

Marine ecosystems are complex, highly diverse, and productive, providing renewable resources to a growing human population. At the same time, the oceans are particularly sensitive to and impacted by anthropogenic stressors (Antão et al., 2020). As such, the scientific community strives to deliver up-to-date information about the state of marine ecosystems so that management decisions are well-informed. Ideally, such decisions use ecosystem-based management (EBM) approaches to preserve ecosystem health and productivity while allowing appropriate human use. EBM is especially relevant in densely populated coastal areas. During this period of rapid environmental change, EBM requires researchers to track ecological change and critical events when, and not well after, they oc-

cur. Fortunately, technological developments in observation methods over the last couple of decades have provided ecologists with a range of new tools for obtaining vast amounts of data from marine ecosystems. These include high-end cameras, echo sounders, and hydrophones, combined with various sensors to measure environmental parameters. Researchers can attach such technologies to cabled observatories or static rigs to assess temporal dynamics, or remotely or autonomously operated vehicles to evaluate spatial variability. However, because these technologies can produce an unprecedented amount of data, which has traditionally required manual processing, ecologists may be reluctant to adopt them as an alternative or supplement to traditional sampling techniques. For example, using traditional gear (e.g. nets and traps) to assess the abundance of fish has been an established sampling technique for centuries and is still used today. These methods are efficient for manual data handling and straightforward: as soon as the fish are caught, counted, and the data entered, it can be analysed by the researchers. On the other hand, detecting and counting fish with cameras is less destructive to animals and habitat, offers high-resolution temporal data, allows researchers to observe behaviour of animals and habitat use, and often provides a more representative estimate of species diversity and relative abundance (Bacheler *et al.*, 2017). However, extracting all of this information from videos manually is a laborious task. Thus, automating this step would undoubtedly encourage more fish biologists to use cameras for data collection.

Many diverse fields of research are undergoing rapid change due to advances in the use of artificial intelligence (AI) for data interpretation. AI offers fast and accurate analysis of the large volumes of data collected by sensors, cameras, and other observation technologies. Off-the-shelf algorithms can now, with high precision, find, count, and classify organisms from digital images and real-time video, (Lopez-Guede *et al.*, 2020; Knausgård *et al.*, 2021; Li and Du, 2021) and detect cryptic patterns in noisy images or acoustic data (Weinstein, 2018). An increasing number of marine ecologists embrace this opportunity, yet initiating collaborations across ecological and data science disciplines can be challenging for several reasons. First, transferring the necessary information to start a project between an ecologist and a computer scientist can be a steep learning curve because knowledge barriers and field-specific jargon can cloud otherwise fruitful discussions and halt progression. Second, ecologists unfamiliar with AI may not be aware of the opportunities available to address a particular problem. Before an ecologist approaches an AI expert, they may need to know about the possibilities and limitations of AI, how to prepare and annotate data sets, and what information to provide the computer scientist to identify the best AI method for the task at hand. Meanwhile, before advising on the possibilities, the computer scientist may find it challenging to understand the underlying ecological question, the data and its inherent variability/noisiness, how it is categorized, and what level of accuracy is needed. Thus, substantial investment in the interdisciplinary partnership is required in order to achieve a common understanding.

This paper aims to bridge the gap between marine ecologists and computer scientists to expedite the initial stages of collaboration. To provide common ground, we describe the most popular and suitable AI techniques for ecological data analysis, including technical concepts. AI is a general term referring to any AI technique that can solve a complicated problem (Goodwin, 2020; Russell and Norvig, 2002). We focus on applicable and well-established methods, namely deep neural networks (DNNs), synonymous with "deep learning" (DL), and learning with supervision (supervised machine learning). Supervised learning requires algorithms to be presented with datasets that have been labelled with accurate information on the region of interest, for example the presence or location of known species, objects, or sound. The algorithms learn to associate the labels with the examples (Christin *et al.*, 2019). With enough training material, the algorithms can produce models that automatically recognize and identify new and unseen examples in other datasets without the need for new labels (LeCun *et al.*, 2015). One of the biggest challenges for supervised learning is the demand for a large, labelled training dataset of sufficient quality to achieve high accuracy (Malde *et al.*, 2020; Beyan and Browman, 2020). Close collaboration between ecologists and computer scientists would likely facilitate and accelerate the dedicated effort required to collect and label representative datasets (Weinstein, 2018; Schneider *et al.*, 2019; Beyan and Browman, 2020).

This paper is organized as follows: In "A non-comprehensive review of DL", we summarize popular DL tools relevant for ecologists and explain standard AI terms. We then provide an overview of machine learning approaches as applied to a series of marine ecology case studies (Table 1). The section "Established cases: identification and quantification of marine biodiversity" describes three cases where AI has been applied to ecological data, namely: fish detection, classification, and tracking in underwater videos; image-based analysis for plankton monitoring; and acoustic monitoring of whales. These applications are generally focused on species and higher-order taxonomic classification for biomonitoring purposes. Yet, DL in ecology research is not limited to these cases and we are confident that the DL toolset will further impact emerging research areas in marine ecology at additional levels of biological organization. Therefore, the section "Emerging cases" continues with four case studies where we see the potential for DL to make a major impact. At the individual level, we show the potential for DL to enable individual visual re-identification of fish using unique patterns (similar to facial recognition) and analysis of fish vocal communication to identify individuals (i.e. vocal recognition) to better understand mating behaviour. At the ecosystem level, we show how DL can aid in ghost fishing gear detection and determining the ecological functions of fish in the carbon cycle. We conclude by discussing technological advances, complexity in data, and acceleration of data collection and labelling through open-source approaches.

A non-comprehensive review of DL

AI is a broad concept, but the most commonly applied technique is machine learning. Machine learning is a set of algorithms that learn from an environment containing data such as images. The most common AI approach used in biology is supervised learning, which is when the data are labelled or categorized so that the algorithms can learn from the data. Conversely, unsupervised learning is when algorithms do not use labelled data but, instead, learn data structures that are reinforced when the algorithms continuously interact with an environment, such as playing a board game. Figure 1 illustrates the overall procedure for training and application of AI with supervised learning.

Among the most popular and widely used AI algorithms are the family of artificial neural networks. A neural network is a set of human brain-inspired networks with artificial neurons and synapses that are trained to approximate an external function, typically mapping from input data (e.g. images) to labelled values or categories (e.g. classes). A neural network consists of a layer of input

Table 1. Machine learning approaches to ecological data applied (green) or explored (blue) in the case studies (C1–C7), and some alternatives (orange). Grey cells indicate no added benefit to using that approach for the task. Approaches: (Ap A): One label per region of interest, (Ap B): One label per image, (Ap C): Pixel-wise segmentation, (Ap D): Ground truth spectrograms with labelled region of interest, (Ap E): Labelled spectrograms with regions of interest, and (Ap F): Segmented time series data.

| Case studies	Object detection		Classification		Segmentation	
	When to use	Possible method	When to use	Possible method	When to use	Possible method
Fish and species counting (C1)	Images with 1+ fish	YOLO with Ap A	Images with 1/0 fish/species	Squeeze-and-excitation with Ap B	Outlines of 1+ regions wanted	R-CNN with Ap C
Plankton analysis (C2)	Images with 1+ organisms	YOLO with Ap A	Images with single organisms	CNN with Ap B	Images with single organism (morphology)	R-CNN with Ap C
Marine bioacoustics (C3)	Spectrograms with 1+ calls	R-CNN with Ap D	Spectrograms with 1/0 calls	CNN with Ap F	Separation for 0+ calls in time series	RNN with Ap E or transformer with Ap E
Re-identification in fish populations (C4)	Images with 1+ fish	YOLO with Ap A	Images with 1/0 individuals	CNN with Ap B	Images with fish outlined	R-CNN with Ap C
Fish vocal communication (C5)	Spectrograms with 1+ individual calls	R-CNN with Ap D	Spectrograms with 1/0 calls	CNN with Ap F	Separation for 0+ calls in time series	RNN with Ap E or transformer with Ap E
Ghost fishing gear detection (C6)	Images with 1+ gear	R-CNN with Ap A	Images with 1/0 gear	CNN with Ap B	Areas with partially dissolved fishing nets	R-CNN with Ap C
Carbon cycling by fish (C7)	Images with 1+ life processes	R-CNN with Ap A	Images with 1/0 life processes	CNN with Ap B	Images or video with moving processes	YOLO with Ap C

neurons connected to the input data and a layer of output neurons mapping to the values or categories to be predicted. It is common to have layers between the input and output, which are referred to as hidden layers. When a network has more than one hidden layer, it is referred to as DL or a DNN.

Neural networks, especially DL, are the go-to machine learning approach for categorizing and recognizing images and sound data. These techniques have won numerous pattern recognition and machine learning competitions for image and sound analytics (Schmidhuber, 2015; Tessler *et al.*, 2017). In recent years, DL has become the predominant analytical technology in many domains, including health (Esteva *et al.*, 2019), customer evaluation (Lessmann *et al.*, 2019), and crisis management (Ben Lazreg *et al.*, 2019; Ben Lazreg, Noori, Comes and Goodwin, 2019). Aquatic ecology has experienced the early stages of the same shift, where object detection and semantic segmentation are being used to identify and locate marine species in raw images, videos, and audio recordings for the purpose of species (Knausgård *et al.*, 2021) and individual (Bogucki *et al.*, 2019) classification, and to quantify abundance. Despite the domination of deeper over more shallow neural networks, there is no need to employ DL models exclusively. Depending on the complexity and the nature of the problem, various models with different depths can be utilized. For example, Kohonen networks, which consist of only one layer, are shallow but useful for biology-related classifications and visualization (Suryanarayana *et al.*, 2008). In addition to identifying and counting fish and other marine animals, there is enormous potential to apply DL to a wide range of data in coastal ecology (Grasso *et al.*, 2019; Marre *et al.*, 2020). In the following subsections, we will briefly go through the basics of DNN. A glossary of AI terms is summarized in Table 2.

DNNs

All neural networks are function approximators; they mimic the function presented in the training data and adapt to this function through an optimization process. During training, the neural networks' weights, which are many real-valued and connected neurons followed by activations, are updated to match the training data. In more detail, the real-valued difference between the predicted output, $\hat{Y}$, and the expected output, Y, is referred to as the loss, which guides the training. For example, Y can be a list of image categories where each value in the vector relates a category to an image, and $\hat{Y}$ is then the neural network's predicted image categories. If the neural network is able to correctly predict image categories, $\hat{Y}$ will be identical to Y and the loss will be zero. The goal of the training process is, generally speaking, to minimize the loss. However, the loss minimization should be done with care since a small loss may indicate that DL has learned specific patterns for each example rather than general trends in the data (i.e. overfitting). To check for overfitting, a separate validation data set is normally employed to independently evaluate the algorithm's performance.

A properly trained network has active or inactive neurons that jointly match the training data and minimize the loss. This is analogous to a series of virtual dials that can be turned completely on, completely off, or somewhere in between, indicating the relevance for each feature. During training, the loss for each neuron is propagated backward through the network so that each neuron's contribution matches the product of the weight and a hyperparameterized learning rate. Hence, each neuron's influence of the loss is matched with a corresponding adjustment of weights, and its adjustment is kept small by the learning rate. When the loss is prop-
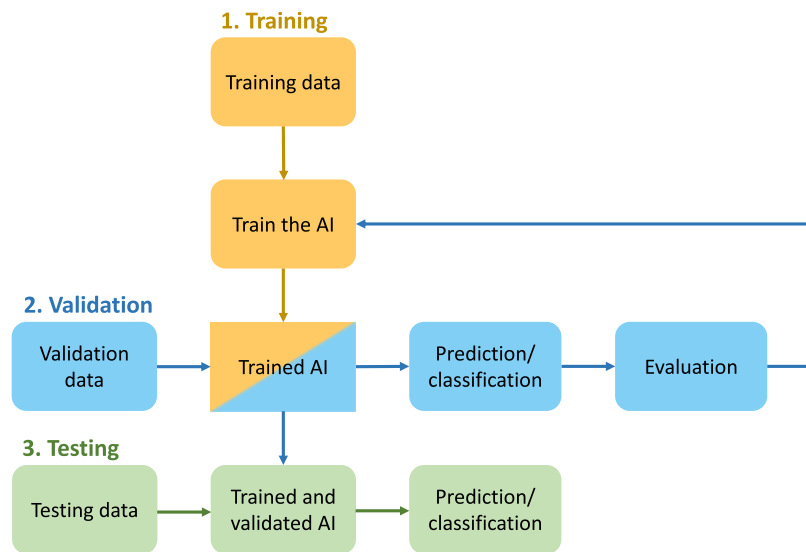
Figure 1. The workflow of AI based strategies. (1) The (yellow) column illustrates the training phase, in which labelled data is used to train the AI algorithm. (2) The first row (blue) shows that the performance of the trained AI is evaluated using a validation data set and the AI algorithm may be updated and refined in this process. (3) The bottom row (green) shows the application phase, using the AI on a test data set once the training and validation are completed.

agated backwards, the dials are turned slightly in the direction that decreases the loss.

A neural network is considered shallow if it has one layer of input neurons, one layer of hidden neurons, and one output layer. The same network would be considered deep if it had more than one hidden layer, and very deep if it had more than ten hidden layers. Any neuron that is not at the input layer combines a weighted sum from active neurons in the previous layer. The sum is then followed by an activation function for the next layer of neurons. Despite popular belief, the depth of the DL may not be proportional to the difficulty of the problem that it can solve. It is not always true that deeper networks solve more complicated issues than shallower networks. Some problems can be solved with shallow networks, but in many cases very deep models empirically outperform the shallow ones for image and sound categorizations. For example, a type of neural network called Residual Networks (sometimes abbreviated to ResNets) often has 18, 34, 50, or 101 layers. Usually, the deeper networks perform better image classification, but occasionally the most shallow network, with 18 layers, is sufficient and even more accurate than the deeper networks (Aloysius and Geetha, 2017).

A poorly trained network is said to 'overfit' when it performs significantly better on the training data compared to the testing data and increased training improves the training results but, at the same time, worsens the testing results. Hence, overfitting is observable by increased training accuracy and decreased testing accuracy. A potential mitigation is to increase the complexity of the network or increase the amount of training data.

A notable limitation of DL is its dependency on vast amounts of training data. The data requirement typically becomes a significant problem in supervised learning, as a successful application in most cases depends on large quantities of human-classified training examples. This challenge is extensively presented in the marine biology domain, as the limited capacity of trained experts makes extensive and quality-assured labelled training databases hard to acquire. A beneficial property of deep unsupervised learning is its

independence of labelled data. However, due to the unsupervised nature, the application area is rather limited in the marine biology domain and has mostly been confined to finding anomalies through re-identification (Dargan *et al.*, 2019; Ferreira *et al.*, 2020a) and data clustering.

Deep semi-supervised learning has emerged in recent years to mitigate the limitations of supervised and unsupervised learning. Semi-supervised approaches combine training on a small amount of labelled data with a subsequent training phase using large amounts of unlabelled data. In applications where there is often a lack of human-classified training data, semi-supervised learning is especially useful.

In the paragraphs below, we summarize typical problems relevant to marine ecology where DNN can be utilized as a promising solution.

Image classification

DNN is the *de facto* standard for machine vision, such as the categorization of images and video files. The most prominent approach among various DNNs is Convolutional Neural Networks (CNNs), which extract relevant features of an image for subsequent classification by a neural network through a series of two-dimensional mathematical convolutional operations with learnable filters of typical sizes 3×3, 5×5, and 7×7 applied in the image pixels. A CNN trained for classification of images finds the function that best maps the input of pixels to a class, e.g. presence of a fish, plankton, or a rope in the photo (Figure 2). Note that the CNN generates small image blocks from the convolutionals of overlapped data within each image. CNN categorizes the image but does not output in which part of the image the object is located.

The first popularized CNN models were LeNet-1–LeNet-5 (Le-Cun *et al.*, 1995), which contain all the basic building blocks still used today. A major advancement, in terms of both architecture and performance, came in 2012 with AlexNet (Krizhevsky *et al.*, 2012). AlexNet achieved an error rate of 15%, which was better than all

Table 2. Glossary table.

Glossary	
Accuracy	Fraction of correct classifications
Activation	A non-linear mathematical operation. It is often used to approximate "turning on" or "turning off" an artificial neuron
Area Under the Curve (AUC)	A summary of the ROC curve that shows capacity of a supervised learning algorithm to distinguish between classes. A perfectly performing algorithm will have an AUC of 1
Attentions	A DL technique to learn and indicate which sequence in a time series or which region of image to pay attention to
Classification	Categorization of input data into classes
Convolution	Mathematical operation that expresses the amount of overlap of one function as it is shifted over another function
Convolutional neural network	A neural network with convolutions, typically used for image classification
DL/DNN	A neural network with more than one hidden layer
Encode-decoder	A neural network that encodes the input data into an internal representation, followed by a neural network that decodes the internal representation, typically to a human readable format
False negative rate	The rate of wrongly predicted negative values
False positive rate	The rate of wrongly predicted positive values
Feature extraction	An operation to select and extract values into feature, typically from unprocessed data
Features	Valued characteristics, typically numeric or structural, representing the input data
Hidden layer	Any layer of neurons in between the input and the output layers
Hyper parameters	User-controlled parameters that influence the model such as number of layers
Layer	A set of neurons that takes data as input and typically does a combination of linear (synapses) and non-linear operations (activation)
Loss	A real number indicating the incorrectness of a single prediction and is typically used to adjust the weights of the neural network
Machine learning	Trainable computer programs that learn the representation of data with an aim to predict never-before-seen data
Model	A representation of what a machine learning program has learned. In a neural network, the model is a combined structure consisting of the network and learned weights of the algorithm
Neural network	A brain-inspired machine learning technique with an input layer (features), one or more hidden layers, and an output layer (predictions)
Neuron	A node that combines input data with learned weights and provides a single output
Object detection	Recognize the presence of an object instance in a location or area
Overfitting	When a model closely predicts the training data but fails to fit testing data
Weights	Real values in a neural network in which each parameter individually prioritizes each data value, and that are updated in the learning process
Pattern	Common trends and regularities in the data such as statistical trends often unique for one category
Pattern recognition	Methods to detect patterns in input data
Precision	The frequency of true positives among all positive predictions
Receiver operating characteristic curve (ROC)	A graph displaying a supervised algorithm's performance at all classification thresholds. Typically, the relationship between the rate of true predictions and the rate of false predictions
Recall	The frequency of correctly identified positive values from all positive values in a data set
Recurrent network	Neural networks that connect between nodes to form a directed graph to detect patterns that occur, often over a time series
Semantic segmentation	The process of partitioning images into labelled regions
Supervised learning	Machine learning that maps an input to a specific, often labelled, output
Synapses	Learned weights on the input data for a layer, i.e. how to prioritize the input features
Labelled training data	Data used for training the model. It is kept separate from testing and validation data
Testing data	Data used for independently evaluating the trained model. It is kept separate from training and validation data
True predictions	Model output that corresponds with the correct values
Underfitting	When a model has not reliably learned the patterns of the data
Unsupervised learning	Machine learning that finds patterns in unlabelled data
Validation data	Data used for verifying the model and tuning the hyper parameters during training

non-neural network architectures, for which the previous best error rate was 26%. These early models suffered from vanishing gradients, meaning that the input data was gradually lost when additional layers were added. This limitation hindered the development of DNN and the performance of the DL models suffered. Later, major innovations included: (1) inception networks (Szegedy *et al.*, 2015), which utilized parallel convolutions of different sizes, (2) residual architecture (He *et al.*, 2016), which added skip connections to allow for an image to both be processed by convolutions and skipped through the network, and (3) squeeze-and-excitation networks (Hu *et al.*, 2018), which introduced a method to add additional parameters to each convolutional block so that the model could adjust the weight of each block. Each of these innovations has enabled larger, more complex networks. Therefore, the rapid advancement of the image classification field indicates that the newest techniques are, in general, much better than earlier ones. Unless there is specific evidence to the contrary, practitioners are advised to choose a more recent approach.
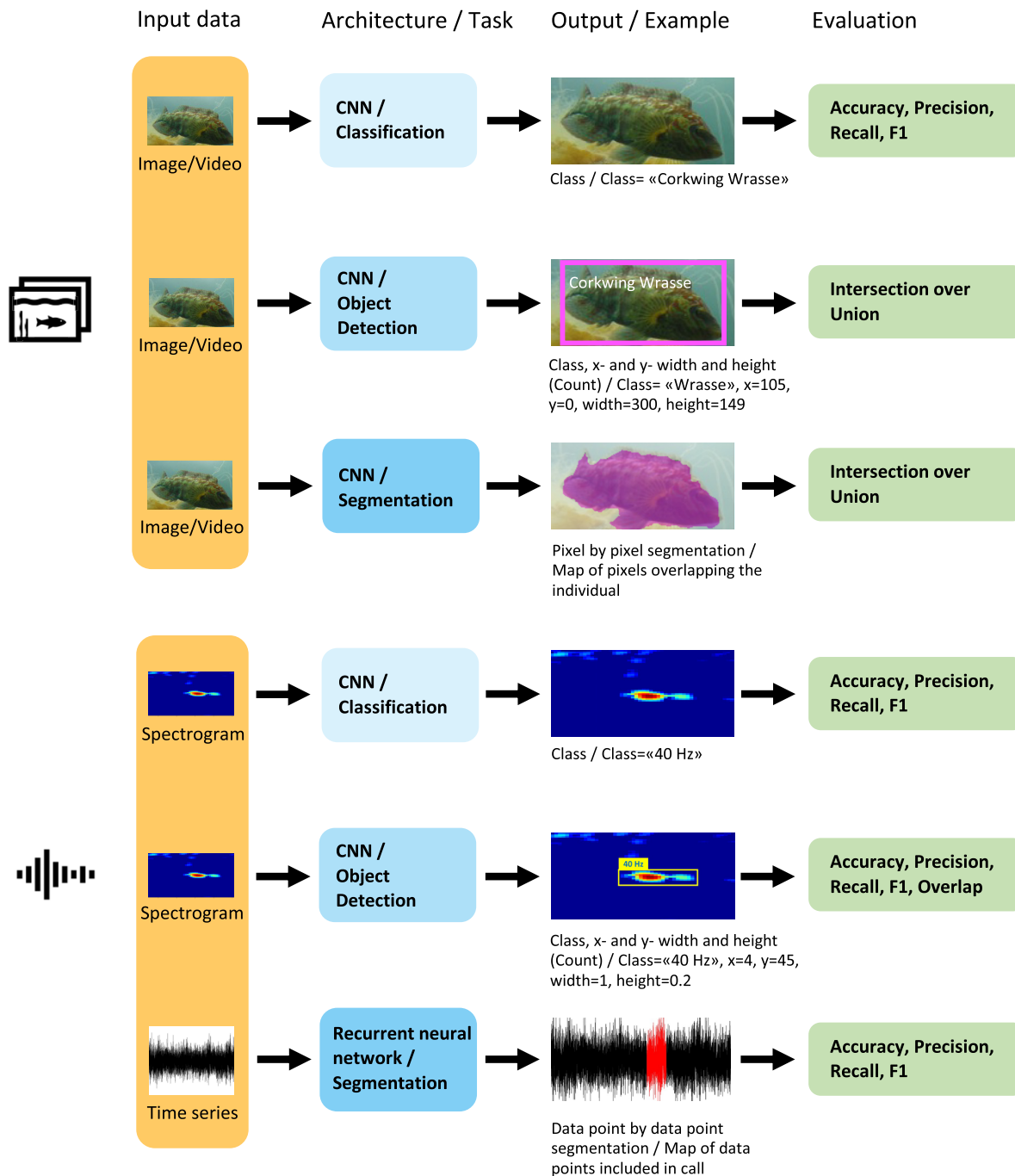
Figure 2. Examples of classification, object detection, and pixel-wise segmentation with illustrations of the techniques applied to fish images or audio files.

Object detection and semantic segmentation

Object detection extends CNN models by detecting regions of interest in the image (Figure 2). In addition to classification, a network trained for object detection can output the *x*- and *y*-location, width, and height of the object of interest. This information is then used to draw a boundary box around the object to be classified, e.g. a fish. In this way, a single image can be divided into multiple regions by generating several boundary boxes, allowing for many classes to be classified within a single image. In practice, this means that we can detect and count objects in an image or a video, e.g. the number of fish. The approach has been extended even further by pixel-wise detection and classification of the entire image. This approach scales down the image with convolutions and pooling operations, followed by reverse order scaling-up of the same image. This is known as an encoder–decoder architecture (Girshick *et al.*, 2014) and allows for categorization of every region in the image at a high level of detail. A commonly used method for object detection is You Only Look Once (YOLO; Redmon *et al.*, 2016; Yu *et al.*,

2021). Variants of Region-based CNN (R-CNN), including Fast R-CNN (Girshick, 2015), and Mask R-CNN (He *et al.*, 2017), are used for pixel-wise segmentation.

Individual identification

A Siamese Neural Network (SNN; Koch *et al.*, 2015) is a type of DL model that contains two identical sub-networks with the same layers, hyper parameters, and weights. The neuron weight updates are mirrored and so can be used to find the similarity of the inputs by comparing vector features. An SNN allows us to detect if two images are the same, e.g. two faces are of the same person or two fish photos are of the same fish taken at a different time. Hence, an SNN can classify a new class without re-training the entire network. Other features include robustness to class imbalance (i.e. data is unequally distributed between classes) and learning efficiency in the semantic similarities between images. However, SNNs need more training data and longer computational time than competing networks.

When training an SNN, a typical loss function used to detect differences in input is a so-called triplet loss, in which the baseline input is compared both with a positive and a negative example. A perfectly trained SNN should have a zero loss for the positive example and a loss for the negative example. For example, when detecting individual marine animals, a comparison between pictures of the same animal should have a smaller loss than a comparison between pictures of two different individuals of the same species. This approach can be used as a method for classification to identify if two pictures include the same individual and verify whether an image consists of an individual that is not part of the training data. Figure 2 provides examples of classification, object detection, and segmentation, and how they are typically evaluated.

Audio signal classification

Audio signal classification is a classic yet challenging field of audio signal processing. In brief, it comprises capturing appropriate features from an audio sequence and employing these features to distinguish the class that the sequence is most likely to fit. Depending on the application domain, one may predict a global signal class with a unique label or a subset of the possible classes with multiple labels. Traditionally, finding appropriate features and designing a suitable classifier are configured as separate procedures. This approach has several drawbacks. The extracted features might not be optimal for the classification objective. Further, certain features may require prior human knowledge, be difficult to describe precisely, or be subjective and unstable. As mitigation, DNN-based approaches are developed to perform feature extraction jointly with classification.

In contrast to feed-forward neural networks, recurrent neural networks (RNNs) contain feedback loops. These loops allow RNNs to use their reasoning from previous data to influence upcoming data, hence lending themselves to process a series of data. This feature is useful when working with data that changes over time, so-called time series, including audio signals. RNNs vary in complexity, from standard RNNs, often called vanilla RNNs, to models with more complex memory elements, including gated recurrent units (GRUs) and long short-term memory (LSTM). The more complex a module is, the more likely it is to learn intricate patterns in the data. However, increasing the complexity also increases training time and

the chance of overfitting. Therefore, the choice of RNN will depend heavily on the level of complexity of the pattern(s) of interest.

A recent trend is to combine RNNs with new variants of feed-forward methods such as attentions (Phan *et al.*, 2019; Chaudhari *et al.*, 2021), combine them with multiple attentions coupled together into the collective concept of transformers (Moritz *et al.*, 2020; Tay *et al.*, 2020), or avoid RNNs altogether and only use transformers. Attention and transformers weigh the significance of input data and, like RNNs, they are designed to handle sequential input data such as audio signals and other time series. However, unlike RNNs, there is no feedback loop, which means that it learns the context for any position in the input sequence. Despite similarities, attentions and transformers are technically not RNNs since they do not rely on recurrent feedback loops but rather a more straightforward feed forward mechanism. However, a transformer has more parameters to learn and is, therefore, more computationally intensive. In practice, developers must balance the complexity of the models with available training time and resources. RNNs, CNNs, and attention modules can be combined to improve the performance of the system. For example, an attention-based convolutional RNN model is utilized for environmental sound classification (Zhang *et al.*, 2020).

Another approach for audio classification is to convert the audio into a visual representation and use image classification as described above. Spectrograms are such a visual representation and can be classified using a DNN or CNN. These mechanisms, alone or in combination, can be utilized for audio classification tasks in marine ecology-related applications. Figure 2 includes examples of classification, object detection, and data point segmentation with CNN and RNN networks for audio categorization.

Evaluation criteria

To evaluate the performance of a trained model, different parameters are utilized by the different approaches, such as accuracy, precision, and recall (Figure 3). Accuracy is the ratio of correct classifications to the total number of classifications. Precision for the positive predictions is the ratio of true positive predictions over the sum of true positive and false positive predictions. The same concept applies to the precision of negative prediction. Recall is the ratio of true positive predictions over the sum of true positive and false negative predictions. A result of a DL algorithm may be precise but not accurate when results are biased but with small variance. A DL algorithm is considered valid if it is both accurate and precise.

For example, if the expected output Y is five images of cod and five images of trout, and the predicted output $\hat{Y}$ correctly identifies all cod and only four of the trout, with one trout wrongly identified as cod, the algorithm is correct nine out of ten times, yielding an overall accuracy of 90%. In this example, the precision for trout is 100%, i.e. all trout were predicted as trout, but only $\frac{5}{6} = 83\%$ for cod, i.e. for all the predicted cod, only 83% are actually cod. The recall for cod will be 100%, i.e. the algorithm identifies all cod, but the recall for trout will be $\frac{4}{5} = 80\%$, i.e. the algorithm only identifies 80% of the trout. Table 3 illustrates this example as a Confusion Matrix.

The parameter used for performance evaluation depends on the data. Accuracy is most suitable if the data set is balanced, meaning an approximately equal number of examples in each class, and where false positives and false negatives have similar implications. But if the data set is imbalanced, which is typical for ecological

Accuracy	Predictions/ Classifications	$\dfrac{\text{Correct}}{\text{Correct + Incorrect}}$
Precision	Predictions/ Classifications	$\dfrac{\text{True Positive}}{\text{True Positive + False Positive}}$
Recall	Predictions/ Classifications	$\dfrac{\text{True Positive}}{\text{True Positive + False Negative}}$
F1	Predictions/ Classifications	$\dfrac{2 * \text{True Positive}}{\text{True Positive + 0.5 (False Positive + False Negative)}}$
IoU	Object Detections/ Segmentations	$\dfrac{\text{Pixel Overlap}}{\text{Pixel Union}}$

Figure 3. Evaluation metrics accuracy, precision, recall, F1-score, and Intersection over Union (IoU) for classifications, predictions, object detections, and segmentations.

Table 3. Example of a Confusion Matrix with five cod and five trout.

		True		
		Cod	**Trout**	
Predicted	Cod	5	0	Recall = 5/(5 + 0) = 100%
	Trout	1	5	Recall = 4/(4 + 1) = 80%
	Pre.	Precision = 5/(5 + 1) = 83%	Precision = 4/(4 + 0) = 100%	Accuracy= 9/10 = 90%

data (e.g. some species are more common than others), precision or recall are better. A high precision relates to a low false-positive rate, whereas a high recall relates to how well the model detects the class in the total data set.

Closely related to precision and recall is the Receiver Operated Characteristics (ROC) curve, where the true positive rate is on the y axis and the false positive rate is on the *x*-axis. Each AI output includes a score that represents certainty. Changing the threshold of acceptable scores affects how conservative the output will be. Only accepting output with a high classification score will result in few false positives but many false negatives. Conversely, accepting output with a lower score will result in more false positives but fewer false negatives. The normalized Area Under the ROC Curve (AUC) describes how well the algorithm works across this range of score thresholds. The value of AUC is a number between 0 and 1, with the latter describing a perfect network.

The F1 score, a unified metric, is a weighted average of precision and recall and, therefore, encompasses both the false positives and false negatives. A general rule of thumb is to use the F1 score for evaluation when unsure based on the other metrics.

For object detection and semantic segmentation, the evaluation should depict how much pixel-wise overlap there is between the predicted and actual objects. The metric used is called 'intersection over union' (IoU). Using fish identification as an example, an IoU of 0 means that there is no overlap between the areas of the predicted fish and actual areas of the fish. Conversely, an IoU of 1 means a perfect pixel-by-pixel overlap between the predicted and actual areas of the fish.

Data

There is no universally right answer as to how much data is needed—generally, the more data, the better. Learning an intricate pattern requires more data than learning a simpler one. For example, for a DL to classify an image as either a sea trout or another fish species with clear morphological differences, such as a cod, it may achieve a near-perfect separator with relatively few samples. However, more data are likely to be required for a model to learn to distinguish sea trout from a closely related species with similarities in appearance, such as salmon, simply because that is a more complex task to learn.

Mitigation for the lack of data means using an existing model with weights pre-trained using other data sources, such as the ImageNet database (Deng *et al.*, 2009a). The typical approach is to first train with an available, sizeable dataset and subsequently train with a smaller but more relevant dataset. In this way, the learning algorithms find the general image patterns from a big dataset (e.g. shapes, species patterns, and face patterns) and the individual differences from the smaller dataset. This process, known as transfer learning, allows researchers to use readily available large data sets like ImageNet to be used on data that seems highly unrelated to the data set of interest. For example, ImageNet has been annotated in categories like 'balloon,' 'tiger,' and 'cat,' yet can be used to train a network to classify fish vocalizations in the Mexican Gulf (Waddell *et al.*, 2021). However, transfer learning has a greater advantage when the domain difference between the data is small. Because distribution in real-world ecological data sets is particularly prominent, undesirable variations can result in misclassifications.

For a classification or object detection task, the dataset needs to be labelled (sometimes referred to as annotated), usually by a human expert (e.g. an ecologist). The labelled data is often referred to as the Y vector. An accurate classifier algorithm should correctly map the input, known as the X vectors (e.g. images) to the appropriate Y vector (the labels). These predicted labels are often referred to as the $\hat{Y}$ vector, regardless of whether the predictions are correct ($\hat{Y}$ matches Y), or incorrect ($\hat{Y}$ does not match Y).

The labels for a classification task are distinct for each input variable, such as a species of fish for each image. This requires manual categorization and labelling of a large set of images. For object detection and semantic segmentation, the labels must also indicate where in the image the object of interest is located. In the case of audio input for RNN and CNN classification, the start and stop times of all events of interest must be labelled in order to segment the data into relevant categories. If object detection is used on spectrograms of audio, the frequency bands must also be labelled, encasing the contours of interest in the spectrogram. As is the case with images, existing labelled datasets also exist for audio, which can be used for training when data is otherwise limited (e.g. the DCLDE 2015 data set for baleen whale social calls; Huang *et al.*, 2016).

A labelled dataset is divided into three separate datasets, as illustrated in Figure 1: training, validation, and testing. The training set is used to train the model, meaning that it tries to find an approach to map the training set's input vector X (images) with the training set's correct labels Y. The validation set follows and is first used to check whether the algorithm can map the validation set's input vector X with the validation set's correct labels Y, which is separate from the training set vectors. This validation set then provides a prediction/classification that can be used to evaluate the performance of the model. After this evaluation, further fine-tuning of the model hyperparameters can be done and the model retrained with the training set if needed (Figure 1). Note that because the model has used the validation set as part of the training process, a new set is needed for the final check of how well the algorithm can classify. This is called the test set. Hence, the training data set is used for the actual optimization process, while the validation data set is used for performance feedback after each training step (epoch). This means that the validation data steers the tuning of hyperparameters and will therefore heavily influence the weights of the final model. In contrast, the testing data is used for verification of the final model only and not for the optimization and selection process while training. Because the test set is kept out of the entire training process, it serves as an independent verification of the resulting model.

If the trained network performs poorly, a possible cause is that the dataset is too small. A simple mitigation is data augmentation, which is used to artificially enlarge the data set and essentially duplicate the training data with modifications. It is important to note that only the training data should be augmented and any augmented image should not have a counterpart in the validation or test sets. Such an inappropriate use of augmentation would falsely increase performance. Typical augmentation techniques include flipping horizontally and vertically, rotating, scaling, cropping, translating the x- and y-coordinate systems, and adding noise. More advanced techniques include using Generative Adversarial Networks for generating new images.

Established cases: identification and quantification of marine biodiversity

The application of DNN provides an alternative to laborious or repetitive manual tasks, such as processing data from underwater

recording equipment. The following section presents three cases in ecological research where DL is already used to alleviate data processing and is likely to become the method of choice. These cases exemplify the methods described in the section "A non-comprehensive review of DL":

Case 1: detection, classification, and tracking of fish in images and videos

Monitoring of fish populations and communities is a central activity within marine management and conservation. Traditional sampling methods to track population trends, estimate abundance, and to infer movement patterns of fish have relied on studies that involve animal handling (i.e. fishing gears, individual tags, and biologgers). These methods are not only invasive, but also time consuming. Developing and applying passive ways to both obtain the necessary data and to speed up analysis are therefore imperative. Today, automated detection, classification, and tracking of small-scale movements of fish through images and video are made possible with DL, an application well-suited to this task.

When selecting AI approaches for monitoring, consider that a real-life underwater scenario typically involves multiple fish present in the same image, which precludes the use of standard classification techniques. A solution to this problem is to introduce object detection before classification. The object detection step discriminates between individuals within an image and separates them, and in this way, prepares the image data for classification. Object detection and classification can be two completely separate steps in a pipeline (Knausgård *et al.*, 2021; Connolly *et al.*, 2021), or integrated as part of an object detector, such as YOLOv1-YOLOv5 (Redmon *et al.*, 2016; Bochkovskiy *et al.*, 2020; Jalal *et al.*, 2020; Yang *et al.*, 2021; Shin *et al.*, 2021; Yu *et al.*, 2021).

Detecting and counting species from still images and videos is relatively straightforward using standard DL object detection algorithms, as described in "A non-comprehensive review of DL". However, a challenge with setting up a detection algorithm is that well-established object detection training datasets, such as Coco (Lin *et al.*, 2014) and ImageNet (Deng *et al.*, 2009b), include few or no images within the category of each species of fish and with very little variation in the background. Thus, the applicability of such data sets is limited to the first part of a transfer-learning process, in which object detection in general is learned. To increase the precision of detection for a specific use-case in marine ecology, one must then train the DNN with more relevant images (e.g. of fish in their natural environment). Collecting and labelling relevant image and video data is therefore central to building a high-performance and robust fish detector. Public datasets are currently an integral part of this research, particularly for fish detection and species identification (e.g. Fish4Knowlege; Fisher *et al.*, 2016, datasets of temperate fish species; Knausgård *et al.*, 2021, and across species, location, and depths, as in NOAA fishery datasets; Link *et al.*, 2015, and the OzFish dataset; Ditria *et al.*, 2021). The best performance by AI in species identification (i.e. classification) is achieved with a specialized CNN that only classifies species without detecting at the same time. The squeeze-and-excitation-based CNN presented in (Knausgård *et al.*, 2021) reached classification accuracy of 99.27% on the Fish4Knowledge dataset (Fisher *et al.*, 2016) and 87.74% on a second temperate species dataset.

Marine researchers often collect videos rather than still images and are interested in tracking the same animal across consecutive frames to obtain information on behaviour (e.g. to estimate swimming speed; Beyan *et al.*, 2015), or to ensure that the same fish is not

counted multiple times (Lopez-Marcano *et al.*, 2021). To continuously follow a moving object's position in a video sequence, such as a swimming fish, object tracking can be used. One way of implementing tracking is to use a detection algorithm that feeds another tracking algorithm with position data. When tracking multiple objects (e.g. a school of fish), a track association decision needs to be made for each object (e.g. each individual fish). Thus, a complete tracking system typically consists of a detection algorithm, association of detection with tracks, and the actual tracking algorithm. In practice, tracking commonly involves Kalman filters or other recursive estimators to enable efficient dynamic tracking of objects (Ristic *et al.*, 2004), including specific fish (Barreiros *et al.*, 2021). Another emerging approach is to let DL solve the entire multi-class tracking problem in one step (Ciaparrone *et al.*, 2020). This one-step approach typically results in a more homogeneous system, but with less fine-scale control than when applying well-understood recursive estimators. Further, a fully integrated CNN-tracking approach leaves less room for the user to include *a priori* information on expected fish dynamics and behaviour. A CNN-only approach will, however, completely avoid the meticulous tuning requirement of mathematical models and Kalman filter parameters.

We see DL as an essential building block for automating image and video analysis where the goal is to quantify, classify, and track fish. DL can either be used in a modular pipeline with separate steps for detection (Knausgård *et al.*, 2021), association, and track building, or as a complete solution to a multi-object tracking problem (Jalal *et al.*, 2020; Yang *et al.*, 2021; Shin *et al.*, 2021). As these DL tools are adaptable for use with different ecosystems or species by virtue of the training datasets used, the potential for AI in monitoring is great.

Case 2: image-based analysis for plankton monitoring

Plankton is a highly diverse group with very different morphologies and sizes ranging from submicrons to a few centimeters, or even a few meters (Lombard *et al.*, 2019). Plankton are responsible for about 50% of global primary production (Field *et al.*, 1998) and constitute the base of many marine food webs. Some species serve as bioindicators of ecosystem health, while others can form toxic blooms with adverse impacts on other marine life, including commercially important fish. Therefore, tracking seasonal, interannual, and spatial changes in plankton composition and abundance is central to coastal monitoring. Image-based monitoring is now an established tool in many regions and it generates an ever-increasing volume of plankton images each year. Various AI approaches have been developed to analyse this data and reduce manual processing. Plankton identification and counting are arguably some of the most useful examples of DL in marine biology. The ultimate goal is fully automated plankton classification without human biases (Culverhouse, 2007). This bias is not trivial, as human experts can only achieve 67–83 % self-consistency during a difficult classification task (Culverhouse *et al.*, 2003), although accuracy is much higher (> 90%) when working with natural plankton samples with many taxa which have variable classification difficulty (Luo *et al.*, 2018).

Several systems for image acquisition and AI analysis of plankton are commercially available (Lombard *et al.*, 2019), including *in situ* (e.g. Imaging FlowCytobot, VPR, and IISIS) and those that image samples, fixed or fresh, on research vessels or in the laboratory (e.g. ZooCam and FlowCam). All approaches share the same basic principles: pictures are taken of the sampling volume and the objects

are segmented (i.e. into individual organisms). Each segment is then classified into one of several pre-defined classes, typically taxonomic or functional groups, but living organisms are always separated from non-living particles. Besides the predicted classification, the algorithms can extract object features (e.g. length, width, and equivalent spherical diameter) and, therefore, provide information on plankton community structure and function (e.g. normalized biomass size spectra; Wang *et al.*, 2020). Seasonal and interannual variability in plankton abundance and composition obtained using these image-based DL methods is comparable with traditional microscopy (e.g. FlowCam; Alvarez *et al.*, 2014).

Initial plankton classification models were based on statistical approaches but soon transitioned into machine learning solutions (Luo *et al.*, 2018; Kerr *et al.*, 2020), including algorithms that classified plankton based on object features such as size or edge, for example Support-Vector Machine (SVM) and Random Forest (RF) algorithms (Faillettaz *et al.*, 2016; Fischer *et al.*, 2020). SVM and RF algorithms reach 70–90% accuracy in classification for the most abundant plankton groups, but rare or cryptic species can still be a problem. These classifiers also cannot extract the object features from the raw data and instead require these to be manually defined by ecologists, a cumbersome process. In order to overcome these issues, CNNs are being proposed, such as collaborative CNNs with configurations to deal with class imbalance (e.g. where one type of plankton is much more frequent than another; Kerr *et al.*, 2020) or when the environment dynamically changes (dataset shift) using a supervised quantification scheme (Orenstein *et al.*, 2020). These CNNs achieve state-of-the-art 90% classification accuracy when classifying independent test sets (e.g. 97% accuracy classifying 0.1 million FlowCam images; Kerr *et al.*, 2020), although accuracy decreases with very many diverse images (e.g. 83% accuracy for 52 million zooplankton images from IISIS; Briseño-Avena *et al.*, 2020). Other approaches to improve accuracy of conventional CNNs are through inclusion of context data (e.g. sampling location and time) in the classifier (Ellen *et al.*, 2019), using unsupervised clustering of data (Schroeder *et al.*, 2020), or combining CNNs with SVM classifiers (Cheng *et al.*, 2020).

DL enables a whole new approach to plankton coastal monitoring by (semi-) automatic analysis of samples either *in situ* or in the lab (Wang *et al.*, 2019). DL is used to monitor long-term, seasonal, and spatial changes in taxonomical groups (Briseño-Avena *et al.*, 2020) and size spectra (Yu *et al.*, 2016; Wang *et al.*, 2020), to track plankton that serve as bioindicators of ecosystem health (Uusitalo *et al.*, 2016), or as an early-warning system for harmful algal blooms that impact higher trophic levels and, ultimately, humans (Gorocs *et al.*, 2018; Orenstein *et al.*, 2020). However, DL cannot yet replace a taxonomist for difficult identification tasks (e.g. identification of certain species or life stages of zooplankton or larval fish), and as such are not yet adequate for studies that require high taxonomic resolution. Experts are also required to create training sets and validate the results. However, manual hours can be reduced if training sets and analysis pipelines are made publicly available (Li *et al.*, 2020; Chen, 2021; Schmid *et al.*, 2021), as well as through the creation of global databases and training sets (e.g. Ecotaxa; Picheral *et al.*, 2017). Ultimately, the combination of traditional physical plankton sampling with autonomous platforms that combine image-based data with data from other sources (e.g. genomics, acoustics, and pigments) appears to be the best way forward for coastal plankton monitoring studies (Gorsky *et al.*, 2019; Lombard *et al.*, 2019).

Case 3: passive acoustic monitoring of whales

The use of long-term underwater passive acoustic monitoring (PAM) recording has grown in the last couple of decades to become an indispensable tool for investigating relative population trends and temporal and spatial migration patterns of a wide range of whale species (Wiggins and Hildebrand, 2016).

For many years, the standard procedure for detecting and classifying whale calls from PAM recording has been to retrieve the sound recording, use a software package like Triton (Wiggins and Hildebrand, 2007) to create spectrograms lasting 1–2 min, then have the spectrograms manually scanned for call contours by a trained data analyst. This method is not only highly labor-intensive, as PAM recording can cover months, if not years, but the results are also subjective (Baumgartner and Mussoline, 2011). As many whale calls are highly stereotypical, algorithms like matched filtering (Giannakis and Tsatsanis, 1990) and spectrogram correlation (Mellinger and Clark, 1997) have successfully been developed for automated call detection. However, these methods tend to work poorly on calls with more variability in frequency modulation. Hence, manually scanning spectrograms continues to be used for many call types. The manual procedure of visually scanning spectrograms for known call contours is very similar to the image classification process. Further, sound classification using DL is becoming well established outside of marine bioacoustics (Piczak, 2015; Sharma *et al.*, 2020; Mushtaq *et al.*, 2021), which has led to significant interest in using CNN for automated whale call detection.

Among the whale calls recently being investigated using CNN are those of the beluga whale (*Delphinapterus leucas*) with an AUC of 0.9906 (Zhong *et al.*, 2020), North Atlantic right whale (*Eubalaena glacialis*) with an AUC of 0.902 (Shiu *et al.*, 2020), killer whales (*Orcinus orca*) with an AUC of 0.9523 (Bergler *et al.*, 2019), and sperm whales (*Physeter macrocephalus*) with 99.5% accuracy in detecting sperm whale clicks in 650 spectrograms (Bermant *et al.*, 2019). A drawback of CNN classification without object detection is that it does not relay information about where in the image an object is located. For example, when examining spectrograms where the *x*-axis is the timeline, no information is included about the call's specific time, nor the number of calls, thus the CNN serves as a "presence" identification tool only. A work-around for this issue has been to make the spectrograms very small, covering only a short timeline (e.g. 2 s; Bergler *et al.*, 2019). When creating a spectrogram, there needs to be an overlap between two consecutive spectrograms. Otherwise, a call located at the intersection of two spectrograms might be missed. Using short spectrograms combined with these overlaps can increase the redundant data up to 20% (Bergler *et al.*, 2019) and thereby increase the computational cost at a similar level. Object detection can solve these issues for whale call detection. For example, a custom-made region-based CNN for detecting regions of interest in combination with a transformed pre-trained CNN for further classifying the regions of interest was successfully trained and tested on the highly variable D call emitted by blue whales and 40 Hz calls emitted by fin whales (*Balaenoptera physalus*; Rasmussen and Širović, 2021).

Looking to the future, use of AI generally, and DL specifically, in automated detection of whale calls in PAM recordings will undoubtedly benefit from the recent developments in neural architecture search (NAS) algorithms (Sun *et al.*, 2019). This new technique of automatically developing network architecture from prefabricated blocks will cut down significantly on the work needed to adapt networks to fit specific species and calls, and make CNN more

accessible for whale researchers. A general move from using CNNs to perform image recognition on spectrograms extracted from the PAM to using DL directly on the PAM is also anticipated. This can be done *via* recurrent networks like long short-time memory networks (Hochreiter and Schmidhuber, 1997) or a recently developed type of network called the transformer (Vaswani *et al.*, 2017).

Emerging cases

A common theme of the established cases mentioned above is that they replace tasks currently conducted by humans—where using DL can reduce costs, labour, and sometimes improved accuracy compared to human analysts. However, DL has the capacity to be applied to solve more complex tasks, detecting patterns in visual and acoustic data that are difficult for humans to reliably detect or discriminate. In this section, we illustrate novel research avenues in which we predict DL will be successfully applied in the near future.

Identifying and characterizing individual phenotypes

Case 4: visual re-identification of individuals in wild fish populations

Methods for individual identification are needed to answer many questions in animal behaviour and ecology, such as growth, movement, and survival inferred from capture–recapture studies (Clutton-Brock and Sheldon, 2010). Currently, the most common approach is to mark animals with various physical identifiers to recognize individuals upon re-sight or re-capture, such as leg rings on birds, number scratching or paint on reptiles, or lip tattoos on larger carnivores. In marine and freshwater systems, capture–recapture studies on fish are most often performed using external number tags or radio-frequency identification (RFID) tags (Pine *et al.*, 2003). However, trapping and tagging surveys are often costly, logistically challenging to conduct, and are intrusive to the animals.

A less invasive and more practical way forward for data collection is to use images or videos from wildlife cameras and perform DL image analysis by taking advantage of natural markings that make individuals identifiable (Schneider *et al.*, 2019). Like humans, many animals have unique features about their individual appearance, such as intricate patterns of spots and stripes on the skin, fur, or feathers. A trained computer vision algorithm can distinguish between individuals as different classes, even when the identifying features are highly complex. CNN networks have been trained to recognize individuals (individual re-identification [Re-ID]) from photos of animals across many taxa, including birds (e.g., 93.6% accuracy; Ferreira *et al.*, 2020b), turtles (e.g., 95% accuracy; Carter *et al.*, 2014), and terrestrial and marine mammals (e.g., 92.5% accuracy; Schofield *et al.*, 2019). Many fish species also have solid visual pigmentation; stripes, spots, or mosaic in contrasting colours that can be clearly seen in images and video surveys (dala Corte and Moschetta, 2016; Hau and Sadovy de Mitcheson, 2019; Mucientes *et al.*, 2019), particularly coastal fish like the corkwing wrasse (*Symphodus melops*; Figure 4). Therefore, development of Re-ID has potential to replace physical tagging for individual identification of teleost fish, and would also be of great value for monitoring, as it could be used to assess individual movement, behaviour, and growth. Re-ID could also solve the problem of double counting when individuals re-enter the field of view, thus improving video-based monitoring of abundance (Aguzzi *et al.*, 2015; Campos-Candela *et al.*, 2018; Perry *et al.*, 2018)

Figure 4. Established and emerging cases for DL in marine biology, from individuals, to species, to ecosystems. Data input type icons represent images/video (cases 1, 2, 4, and 6), audio (cases 3 and 5), and large-scale environmental monitoring data that is often stored on remote servers (i.e. "the cloud"; case 7). Photo credits: Geir Eliassen (ghost fishing gear), Frithjof Moy/Havforskningsinstituttet (kelp forest).

As far as we are aware, Re-ID by CNN has not been tested in wild populations. One of the challenges preventing the widespread development of AI-based Re-ID is the need for photos or videos of known individuals, independently validated with high certainty, for the training and validation of the algorithm. One solution to this problem is collecting data by using remote detection systems, such as RFID technology, to identify individuals tagged with passive integrated transponders (PITs). By combining PIT-tagging with RFID and synchronized underwater cameras, a large, automatically labelled dataset of many individuals could be created over a relatively short time (Schneider *et al.*, 2019; Ferreira *et al.*, 2020b).

Case 5: inter- and intra-individual variability in fish vocal communications

Acoustic communication is a fundamental component of animal life, especially for aquatic species for which visual cues are not as effective (Tessler *et al.*, 2017). For example, many fish hear their species mating choruses from several kilometers away (Winn, 1964). Subtle variation in complex acoustic signals is challenging for humans to detect or interpret. Furthermore, using algorithms to detect patterns that defy human perception has technological limitations, including processing high volumes of noisy, real-time acoustic data. Using algorithms to detect acoustic signaling presents the additional challenge of source identification in moving animals.

However, advances both in audio recording technologies and in DL algorithms that can detect and classify acoustic signals in natural settings have opened up new systems for study, both on land and at sea (Parsons *et al.*, 2009). These technologies unlock the potential for understanding inter- and intra-individual variation in acoustic communication of fish.

Marine mammals are relatively well studied in this respect, as vocalizations can be classified at the species, population, and even individual levels (e.g., Case 3). However, understanding of the diversity of fish vocalizations and how these vary within species is poorly understood. Moving beyond species-level to population- and individual-level classification of vocalizations is necessary to understand the ecological and evolutionary consequences of acoustic communication in fish and the potential impacts of anthropogenic noise pollution on them. Further, individual-level classification permits a better understanding of intra-individual variation in communication, which is necessary for understanding the role of vocalizations in fish behaviour and personality.

A prime example is Atlantic cod (*Gadus morhua*), which use drumming vocalizations during social interactions, particularly during mating (Brawn, 1961). Yet, our understanding of inter- and intra-individual variation in drumming is limited. There is potential to catalog individual variation in sound production using DL algorithms (Deng and Yu, 2014). Fine-scale individual variation in fish sounds, especially without *a priori* knowledge, is

beyond human perception. Thus, automating this task requires DL approaches that do not rely on labelled training sets. Specifically, CNNs can detect and classify fish sounds by implementing a transformer network (Deng and Yu, 2014). Transformer networks work solely on optimized self attention (looking at other positions in the input sequence for clues that can help lead to a better encoding of each input element) and are currently state-of-the-art in translation tasks. The transformer network is rapidly replacing RNNs previously used for this kind of task, as it solves two of the problems inherent in RNNs: 1) long computing times due to serial processing and 2) vanishing gradients.

Ecosystem

Case 6: ghost fishing gear detection

When fishing gear is lost, the continued mortality of fish, crustaceans, and other species caught in the gear is termed ghost fishing (Brown and Macfadyen, 2007). The problem is widespread and high rates of fish trap loss are reported (Vadziutsina and Riera, 2020). Using DL to detect and locate lost gear can greatly increase the efficiency of clean-up efforts, as human effort could then focus on retrieving gear (e.g., using remotely operated vehicles). Detection of ghost fishing gear has been achieved using side-scan sonar for data acquisition followed by feature cloud generation, which involves looking for objects in an image by identifying areas of high entropy, then clustering and noise reduction to separate the objects from noise by looking for clusters of the identified areas (Labbe-Morissette and Gauthier, 2020).

The next step is using autonomous object detection to extract the location of lost fishing gear. The detection of lost fishing nets using a towed underwater camera followed by automatic object detection has been achieved with a region-based CNN (R-CNN; Politikos *et al.*, 2021). In that study, fishing nets were detected with higher precision than any other type of marine litter. Detection of more types and features of fishing gear is of interest to researchers and clean-up efforts (e.g. whether the feature detected is a trap, fyke net, or ropes). Image classification may be an effective approach to provide this level of detail, where low resolution images are not usually a hindrance for successful image classification. As well as video, side-scan sonar on autonomously operated vehicles could provide the data needed for this approach. Towed underwater cameras may represent a low-cost option for data collection, whereas autonomously operated vehicles equipped with side-scan sonar represent a high-cost option.

Case 7: carbon cycling by fish

The ocean sinks approximately one-third of greenhouse gas emissions out of the atmosphere, including carbon dioxide. The ocean carbon sink is driven by a physical and a biological pump. As well as plankton and bacteria, fish contribute to the biological pump, with recent estimates suggesting 16 percent of sinking carbon could be due to fish (Saba *et al.*, 2021). However, the role of fish in the biological pump is not well understood (Martin *et al.*, 2021). The data on fish required to improve our understanding relates to metabolic use and excretion of consumed carbon and other nutrients; properties of carbon and nutrient outputs and their fate in the environment; habitat use and connectivity of ecosystems; and physical interactions with extrinsic carbon and nutrients in the environment. As well as advancing knowledge of the role of fish, this knowledge

could inform effective management approaches to maintaining or restoring ecosystem carbon function. As an emerging field, zoogeochemistry has the advantage that much of the relevant data are already published for other purposes. For example, metabolic rates and behavioural data is already published for many commercially important species through fisheries and climate change research. Using AI in this field has the potential to expedite a better understanding of fish ecological functions, effects of human disturbance, and therefore potential management of important carbon sink habitats. Here, we present a few of the options available to apply DL to zoogeochemistry research.

In habitats where visual sampling is possible, video images could be used with object detection, classification, and tracking to identify the presence or absence, behaviour, and features of particles from fish and their short-term fate (e.g., defecation, spawning, and whether material reaches and settles on the sea floor). This could inform estimates of the volume of carbon transferred into or out of a habitat by fish, and the short-term fate of the carbon or nutrient they release. Methods that use AI computer vision to determine the connectivity of fish populations can also be of value in estimating carbon flow (Lopez-Marcano *et al.*, 2021). The long-term fate of carbon and nutrients depends on physical, chemical, and biological conditions of the environment. Graph networks, which map out a physical system using nodes to form a graph, have recently been used to simulate the physical behaviour of materials (Sanchez-Gonzalez *et al.*, 2020). This technology has potential application to estimating the probable fate of carbon and nutrient outputs through simulations that combine oceanographic data with features of the carbon released by fish. With many variables to consider, recent approaches to assessing carbon contained in sediments in different habitats include a combination of survey (acoustic and image-based) and bathymetry data, modelling, and remote ground-truthing (Wilson *et al.*, 2018; Hunt *et al.*, 2020). The current approach is manual, but there is potential for AI application to link habitats to carbon fates and make spatial and temporal estimates on cycling and sinking of carbon and nutrients. Graph networks could be applied to generate probable long-term fates of carbon and nutrient outputs based on isotopes (Lyubchich and Woodland, 2019), or where simulations can be informed from video observations and environmental parameters such as season, temperature, currents, and maps of habitat type (Sanchez-Gonzalez *et al.*, 2020).

As has been mentioned in earlier cases, biological data for fish is partially or fully available for commercially targeted species in online databases (e.g. Fishbase). Such databases have been used to generate estimates of nutrient output from fish, such as nitrogen and phosphorous (Schiettekatte *et al.*, 2020). AI can be trained on these databases to estimate ecological and behavioural carbon flows, including on food webs and habitat use (Bohan *et al.*, 2011). This training could then be applied to generate estimates for species where ecological data is limited, such as deep-sea fish. The research needs for deep-sea fish are urgent as commercial interest is increasing at the same time as the significance of these species in moving carbon from surface waters to the deep sea is beginning to be explored, but data collection methods are expensive, time consuming, and patchy (Martin *et al.*, 2020). In this instance, DL could be used to detect probable carbon flows by using logic-based machine learning (i.e. techniques that incorporate background knowledge or rules; Bohan *et al.*, 2011).

Discussion

We are entering a new era in ocean research and management thanks to new technological developments in observational methods combined with AI-supported data analysis. Data collection, processing, and interpretation are at the core of ecological studies and biodiversity monitoring. Scientists are increasingly relying on indirect observations from various sensors generating large and complex data sets, especially in the aquatic environment. Thus, we envision that within a decade, marine researchers will firmly integrate AI and DL in data collection and analysis within most sub-fields of applied marine biology. This development will only continue to accelerate with new generations of biologists better educated in computer science and informatics (Weinstein, 2018).

Non-human, autonomous, and remote platforms such as cabled observatories, autonomous underwater vehicles or gliders, and ships of opportunity will have a pivotal role in ocean monitoring (Whitt *et al.*, 2020). These platforms will record continuous, real-time information on water physics, chemistry, community composition, and biomass of plankton, fish, and other marine species. For example, long-term monitoring of harmful bloomforming plankton species can be achieved using inexpensive image technology anchored to piers (Gorocs *et al.*, 2018; Orenstein *et al.*, 2020). Similarly, changes in whale population trends and migrations can be investigated using PAM (Szesciorka *et al.*, 2020). These methods are likely to decrease reliance on manual analysis or direct sampling through more invasive, expensive, time-consuming, or labour-intensive traditional approaches. This new way of observing the ocean will generate large volumes of data that will only be feasible to analyse with the help of AI. Therefore, AI will play a key role in making routine processes more time-efficient and alleviate the manual work required. For example, a trained data analyst currently needs 50–350 workdays to manually scan 1 year's worth of PAM recordings for whale calls (Woods and Sirovic, pers. comm.). In contrast, the same task can be accomplished by a trained neural network in approximately 4 workdays (Bergler *et al.*, 2019). Fully automated coastal monitoring systems will be faster and more efficient at detecting changes of interest, such as necessitating warnings to the public where toxic algae are abundant and enabling redirection of boat traffic where whales are moving across shipping routes. Altogether, this monitoring information will be valuable in the development of biological indicators and in integrated assessments to support EBM (Tam *et al.*, 2017).

It is important to emphasize that expert work will always be needed to create and correctly label training sets and revise automated analyses, such as when new species enter a system. A model's accuracy performance is likely to decrease significantly when new species that are not part of the training data are introduced. There is no established automated approach to detect when models need retraining. Repeated validation of the models is required to ensure up-to-date performance. This anticipated demand emphasizes the need to develop multidisciplinary skills in researchers at all career stages, as well as the skills required to form fruitful interdisciplinary collaborations (McDonald *et al.*, 2018).

Collaborative work based on open access and sharing culture (from model configurations to training sets) will be essential to advance this future. While this is a common practice within AI communities, the culture of marine science is not as open. However, funding agencies, publishers, and institutions are increasingly enforcing open access for data generated *via* public funds. The FAIR Principles for scientific data management and stewardship are now widely adopted (Wilkinson *et al.*, 2019). These emphasize improving the access, utility, and reuse of data by machines in addition to individual researchers. As such, they may play a vital role in applying AI to the marine domain. Some collaborative initiatives are underway to create global databases for plankton and benthic images and training sets (e.g., EcoTaxa; Picheral *et al.*, 2017 and BIIGLE; Langenkämper *et al.*, 2017), as well as pipelines (Chen, 2021). Ultimately, we envision libraries of images, videos, and metadata available globally, similarly to the open access GenBank database for sequence information and associated metadata for genetic material hosted by the National Center for Biotechnology Information (NCBI) in the United States.

Conclusions and future directions

We have provided examples of how image and audio analysis are already used to analyse marine biodiversity distribution and dynamics in non-invasive ways, emerging applications of AI, and a look at what the future of AI in marine ecology requires. The United Nations Decade of the Ocean has just started, with the aim of achieving "a healthy, safe, and resilient ocean for sustainable development by 2030 and beyond". We have shown that AI will be key to achieve this goal by developing new technology to uncover new aspects of and potential threats to marine ecosystems' structures and functions, thereby informing EBM. This new knowledge will directly address several of the key challenges identified for the Decade, from effective EBM and biodiversity conservation, to creating a digital representation of the ocean and delivering data, knowledge, and technology to all. The Decade of the Ocean initiative promotes global cooperation and interdisciplinary efforts at all levels, which are at the core of how AI-linked marine studies will progress. Where researchers have the opportunity to gather large amounts of complex ecological data, unfamiliarity with AI jargon and the latest developments should not prevent collaborations with data and computer scientists to support EBM of ocean resources during this time of rapid change.

Funding

Data availability statement

No original data are presented.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

REFERENCES

Aguzzi, J., Doya, C., Tecchio, S., De Leo, F. C., Azzurro, E., Costa, C., Sbragaglia, V. *et al.* 2015. Coastal observatories for monitoring of fish behaviour and their responses to environmental changes. Reviews in Fish Biology and Fisheries, 25: 463–483.

Aloysius, N., and Geetha, M. 2017. A review on deep convolutional neural networks. *In* Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP)', IEEE, pp. 0588–0592.

Alvarez, E., Moyano, M., Lopez-Urrutia, A., Nogueira, E., and Scharek, R. 2014. Routine determination of plankton community composition and size structure: a comparison between FlowCAM and light microscopy. Journal of Plankton Research, 36: 170–184.

Antão, L. H., Bates, A. E., Blowes, S. A., Waldock, C., Supp, S. R., Magurran, A. E., Dornelas, M. *et al.* 2020. Temperature-related biodiversity change across temperate marine and terrestrial systems. Nature Ecology and Evolution, 4: 927–933.

Bacheler, N. M., Geraldi, N. R., Burton, M. L., Muñoz, R. C., and Kellison, G. T. 2017. Comparing relative abundance, lengths, and habitat of temperate reef fish using simultaneous underwater visual census, video, and trap sampling. Marine Ecology Progress Series, 574: 141–155.

Barreiros, M. d. O., Dantas, D. d. O., Silva, L. C. d. O., Ribeiro, S., and Barros, A. K. 2021. Zebrafish tracking using YOLOv2 and Kalman filter. Scientific Reports, 11: 3219.

Baumgartner, M. F., and Mussoline, S. E. 2011. A generalized baleen whale call detection and classification system. The Journal of the Acoustical Society of America, 129: 2889–2902.

Ben Lazreg, M., Goodwin, M., and Granmo, O.-C. 2019. An iterative information retrieval approach from social media in crisis situations. *In* Proceedings of the 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), IEEE, pp. 1–8.

Ben Lazreg, M., Noori, N., Comes, T., and Goodwin, M. 2019. Not a target. a deep learning approach for a warning and decision support system to improve safety and security of humanitarian aid workers. *In* Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 378–382.

Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Nöth, E., Hofer, H. *et al.* 2019. Orca-spot: an automatic killer whale sound detection toolkit using deep learning. Scientific reports, 9: 1–17.

Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S., and Gruber, D. F. 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Scientific Reports 9: 1–10.

Beyan, C., Boom, B. J., Liefhebber, J. M. P., Shao, K.-T., and Fisher, R. B. 2015. Natural swimming speed of *Dascyllus reticulatus* increases with water temperature. ICES Journal of Marine Science, 72: 2506–2511.

Beyan, C., and Browman, H. I. 2020. Setting the stage for the machine intelligence era in marine science. ICES Journal of Marine Science 77: 1267–1273.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. 2020. YOLOv4: optimal speed and accuracy of object detection. arXiv preprint not peer reviewed arXiv:2004.10934.

Bogucki, R., Cygan, M., Khan, C. B., Klimek, M., Milczek, J. K., and Mucha, M. 2019. Applying deep learning to right whale photo identification. Conservation Biology, 33: 676–684.

Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A., and Tamaddoni-Nezhad, A. 2011. Automated discovery of food webs from ecological data using logic-based machine learning. PLoS ONE, 6: e29028.

Brawn, V. M. 1961. Reproductive behaviour of the Cod (Gadus Callarias L.). Behaviour, 18: 177–197.

Briseño-Avena, C., Schmid, M. S., Swieca, K., Sponaugle, S., Brodeur, R. D., and Cowen, R. K. 2020. Three-dimensional cross-shelf zooplankton distributions off the Central Oregon coast during anomalous oceanographic conditions. Progress in Oceanography, 188: 102436. .

Brown, J., and Macfadyen, G. 2007. Ghost fishing in European waters: impacts and management responses. Marine Policy, 31: 488–504.

Campos-Candela, A., Palmer, M., Balle, S., and Alós, J. 2018. A camera-based method for estimating absolute density in animals displaying home range behaviour. Journal of Animal Ecology, 87: 825–837.

Carter, S. J., Bell, I. P., Miller, J. J., and Gash, P. P. 2014. Automated marine turtle photograph identification using artificial neural networks, with application to green turtles. Journal of Experimental Marine Biology and Ecology, 452: 105–110.

Chaudhari, S., Mithal, V., Polatkan, G., and Ramanath, R., 2021. An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology (TIST), 12: 1–32, doi:10.1145/3465055.

Cheng, X., Ren, Y., Cheng, K., Cao, J., and Hao, Q. 2020. Method for training convolutional neural networks for in situ plankton image recognition and classification based on the mechanisms of the human eye. Sensors (Basel), 20: 2592.

Christin, S., Hervet, E., and Lecomte, N. 2019. Applications for deep learning in ecology. Methods in Ecology and Evolution, 10: 1632–1644.

Ciaparrone, G., Sánchez, F. L., Tabik, S., Troiano, L., Tagliaferri, R., and Herrera, F. 2020. Deep learning in video multi-object tracking: a survey. Neurocomputing, 381: 61–88.

Clutton-Brock, T., and Sheldon, B. C. 2010. Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. Trends in Ecology and Evolution, 25: 562–573.

Connolly, R., Fairclough, D., Jinks, E., Ditria, E., Jackson, G., Lopez-Marcano, S., Olds, A. *et al.* 2021. Improved accuracy for automated counting of a fish in baited underwater videos for stock assessment. bioRxiv. doi: 10.1101/2021.02.01.429285.

Culverhouse, P. F. 2007. Natural object categorization: man versus machine. Automated Taxon Identification in Systematics: Theory, Approaches and Applications, MacLeod, CRC Press, pp. 25–46.

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V., and González-Gil, S. 2003. Do experts make mistakes? A comparison of human and machine indentification of dinoflagellates. Marine Ecology Progress Series, 247: 17–25.

dala Corte, R. B., Moschetta, J. B., and Becker, F. G. 2016. Photo-identification as a technique for recognition of individual fish: a test with the freshwater armored catfish *Rineloricaria aequalicuspis* Reis & Cardoso, . . 2001 (Siluriformes: Loricariidae). Neotropical Ichthyology, 01: e150074.

Dargan, S., Kumar, M., Ayyagari, M. R., and Kumar, G. 2019. A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering, 27: 1071–1092.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. 2009a. ImageNet: a large-scale hierarchical image satabase. *In* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 'CVPR09'. IEEE Computer Society.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. 2009b. Imagenet: a large-scale hierarchical image database. *In* Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 248–255.

Deng, L., and Yu, D. 2014. Deep learning: methods and applications. Foundations and Trends in Signal Processing. 7: 197–387.

Ditria, E. M., Connolly, R. M., Jinks, E. L., and Lopez-Marcano, S. 2021. Annotated video footage for automated identification and counting of fish in unconstrained seagrass habitats. Frontiers in Marine Science.8: 629485.

Ellen, J. S., Graff, C. A., and Ohman, M. D. 2019. Improving plankton image classification using context metadata. Limnology and

Oceanography: Methods, 17: 439–461. https://aslopubs.onlinelibrary.wiley.com/doi/full/10.1002/lom3.10234

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C. *et al.* 2019. A guide to deep learning in healthcare. Nature Medicine, 25: 24–29.

Faillettaz, R., Picheral, M., Luo, J. Y., Guigand, C., Cowen, R. K., and Irisson, J.-O. 2016. Imperfect automatic image classification successfully describes plankton distribution patterns. Methods in Oceanography, 15: 60–77.

Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R. *et al.* 2020a. Deep learning-based methods for individual recognition in small birds. Methods in Ecology and Evolution, 11: 1072–1085.

Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R. *et al.* 2020b. Deep learning-based methods for individual recognition in small birds. Methods in Ecology and Evolution, 11: 225300989.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science, 281: 237–240.

Fischer, A. D., Hayashi, K., McGaraghan, A., and Kudela, R. M. 2020. Return of the "age of dinoflagellates" in Monterey Bay: drivers of dinoflagellate dominance examined using automated imaging flow cytometry and long-term time series analysis. Limnology and Oceanography, 65: 2125–2141. .

Fisher, R. B., Chen-Burger, Y.-H., Giordano, D., Hardman, L., Lin, F.-P. *et al.* 2016. Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data, 104, Springer.

Giannakis, G. B., and Tsatsanis, M. K. 1990. Signal detection and classification using matched filtering and higher order statistics. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38: 1284–1296.

Girshick, R. 2015. Fast r-cnn. *In* Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

Goodwin, M. 2020. AI: Myten om maskinene, Humanist forlag.

Gorocs, Z., Tamamitsu, M., Bianco, V., Wolf, P., Roy, S., Shindo, K., Yanny, K. *et al.* 2018. A deep learning-enabled portable imaging flow cytometer for cost-effective, high-throughput, and label-free analysis of natural water samples. Light: Science and Applications, 7: 52312417.

Gorsky, G., Bourdin, G., Lombard, F., Pedrotti, M. L., Audrain, S., Bin, N., Boss, E. *et al.* 2019. Expanding Tara oceans protocols for underway, ecosystemic sampling of the ocean-atmosphere interface during Tara Pacific expedition (2016–2018). Frontiers in Marine Science, 6: 750. https://www.frontiersin.org/article/10.3389/fmars.2019.00750.

Grasso, I., Archer, S. D., Burnell, C., Tupper, B., Rauschenberg, C., Kanwit, K., and Record, N. R. 2019. The hunt for red tides: deep learning algorithm forecasts shellfish toxicity at site scales in coastal maine. Ecosphere, 10: e02960.

Hau, C. Y., and Sadovy de Mitcheson, Y. 2019. A facial recognition tool and legislative changes for improved enforcement of the cites appendix ii listing of the humphead wrasse, *Cheilinus undulatus.* Aquatic Conservation: Marine and Freshwater Ecosystems, 29: 2071–2091.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. 2017. Mask r-cnn. *In* Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. Neural Computation, 9: 1735–1780.

Hu, J., Shen, L., and Sun, G. 2018. Squeeze-and-excitation networks. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Huang, H. C., Joseph, J., Huang, M. J., and Margolina, T. 2016. Automated detection and identification of blue and fin whale foraging calls by combining pattern recognition and machine learning techniques. *In* Proceedings of the OCEANS 2016 MTS/IEEE Monterey. IEEE, pp. 1–7.

Hunt, C., Demšar, U., Dove, D., Smeaton, C., Cooper, R., and Austin, W. E. 2020. Quantifying marine sedimentary carbon: a new spatial analysis approach using seafloor acoustics, imagery, and groundtruthing data in Scotland. Frontiers in Marine Science, 7: 588.

Jalal, A., Salman, A., Mian, A., Shortis, M., and Shafait, F. 2020. Fish detection and species classification in underwater environments using deep learning with temporal information. Ecological Informatics, 57: 101088.

Kerr, T., Clark, J. R., Fileman, E. S., Widdicombe, C. E., and Pugeault, N. 2020. Collaborative deep learning models to handle class imbalance in flowcam plankton imagery. IEEE Access, 8: 170013–170032.

Knausgård, K. M., Wiklund, A., Sørdalen, T. K., Halvorsen, K. T., Kleiven, A. R., Jiao, L., and Goodwin, M. 2021. Temperate fish detection and classification: a deep learning based approach. Applied Intelligence, 1–14.

Koch, G., Zemel, R., and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. *In* Proceedings of the ICML Deep Learning Workshop, 2, Lille, France.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. in Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., eds, Advances in Neural Information Processing Systems, 25, Curran Associates, Inc., pp. 1097–1105. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Labbe-Morissette, G., and Gauthier, S. 2020. Unsupervised extraction of underwater regions of interest in side scan sonar imagery. Journal of Ocean Technology, 15, pp. 96–108.

Langenkämper, D., Zurowietz, M., Schoening, T., and Nattkemper, T. W. 2017. Biigle 2.0 – browsing and annotating large marine image collections. Frontiers in Marine Science, 4: 83.

LeCun, Y., Bengio, Y. *et al.* 1995. Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks, 3361: 1995.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. Nature, 521: 436–444.

Lessmann, S., Haupt, J., Coussement, K., and De Bock, K. W. 2019. Targeting customers for profit: an ensemble learning framework to support marketing decision-making. Information Sciences, 557: 286–301.

Li, D., and Du, L. 2021. Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish. Artificial Intelligence Review, 1: 1–40.

Li, J. Yang, Z. and Chen, T. 2021. 'Dyb-planktonnet'. doi: 10.21227/875n-f104.

Li, Q., Sun, X., Dong, J., Song, S., Zhang, T., Liu, D., Zhang, H. *et al.* 2020. Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. ICES Journal of Marine Science, 77: 1427–1439.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. *et al.* 2014. Microsoft coco: common objects in context. *In* Proceedings of the European Conference on Computer Vision, Springer, pp. 740–755.

Link, J. S., Griffis, R. B., and Busch, D. S. 2015. NOAA fisheries climate science strategy. NOAA Technical Memorandum NMFS-F/SPO-155. U.S.Department of Commerce.

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., Sosik, H. M. *et al.* 2019. Globally consistent quantitative observations of planktonic ecosystems. Frontiers in Marine Science, 6: 196.

Lopez-Guede, L.-V., Marini, F., and Johnsen, A. 2020. Video image enhancement and machine learning pipeline for underwater animal detection and classification at cabled observatories. Sensors, 20: 726.

Lopez-Marcano, S., Brown, C. J., Sievers, M., and Connolly, R. M. 2021. The slow rise of technology: computer vision techniques in fish population connectivity. Aquatic Conservation Marine and Freshwater Ecosystems, 31: 210–217.

Luo, J. Y., Irisson, J.-O., Graham, B., Guigand, C., Sarafraz, A., Mader, C., and Cowen, R. K. 2018. Automated plankton image analysis using convolutional neural networks. Limnology and Oceanography Methods, 16: 814–827.

Lyubchich, V., and Woodland, R. J. 2019. Using isotope composition and other node attributes to predict edges in fish trophic networks. Statistics and Probability Letters, 144: 63–68.

Malde, K., Handegard, N. O., Eikvil, L., and Salberg, A.-B. 2020. Machine intelligence and the data-driven future of marine science. ICES Journal of Marine Science, 77: 1274–1285.

Marre, G., Braga, C. D. A., Ienco, D., Luque, S., Holon, F., and Deter, J. 2020. Deep convolutional neural networks to monitor coralligenous reefs: operationalizing biodiversity and ecological assessment. Ecological Informatics, 59: 101110.

Martin, A. H., Pearson, H. C., Saba, G. K., and Olsen, E. M. 2021. Integral functions of marine vertebrates in the ocean carbon cycle and climate change mitigation. One Earth, 4: 680–693.

Martin, A., Boyd, P., Buesseler, K., Cetinic, I., Claustre, H., Giering, S., Henson, S. *et al.* 2020. The oceans' twilight zone must be studied now, before it is too late. Nature, 580: 26–28.

McDonald, K. S., Hobday, A. J., Fulton, E. A., and Thompson, P. A. 2018. Interdisciplinary knowledge exchange across scales in a globally changing marine environment. Global Change Biology, 24: 3039–3054.

Mellinger, D. K., and Clark, C. W. 1997. Methods for automatic detection of mysticete sounds. Marine and Freshwater Behaviour and Physiology, 29: 163–181.

Moritz, N., Wichern, G., Hori, T., and Le Roux, J. 2020. All-in-one transformer: unifying speech recognition, audio tagging, and event detection. *In* Proceedings of INTERSPEECH. 3112–3116.

Mucientes, G., Irisarri, J., and Villegas-Ríos, D. 2019. Interannual fine-scale site fidelity of male ballan wrasse labrus bergylta revealed by photo-identification and tagging. Journal of Fish Biology, 95: 1151–1155.

Mushtaq, Z., Su, S.-F., and Tran, Q.-V. 2021. Spectral images based environmental sound classification using CNN with meaningful data augmentation. Applied Acoustics, 172: 107581.

Orenstein, E. C., Kenitz, K. M., Roberts, P. L. D., Franks, P. J. S., Jaffe, J. S., and Barton, A. D. 2020. Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers. Limnology and Oceanography, Methods, 18: 739–753. https://aslopubs.onlinelibrary.wiley.com/doi/full/10.1002/lom3.10399

Orenstein, E. C., Ratelle, D., Briseno-Avena, C., Carter, M. L., Franks, P. J. S., Jaffe, J. S., and Roberts, P. L. D. 2020. The scripps plankton camera system: a framework and platform for in situ microscopy. Limnology and Oceanography, Methods, 18: 681–695.

Parsons, M. J., McCauley, R. D., Mackie, M. C., Siwabessy, P., and Duncan, A. J. 2009. Localization of individual mulloway (*Argyrosomus japonicus*) within a spawning aggregation and their behaviour throughout a diel spawning period. ICES Journal of Marine science, 66: 1007–1014.

Perry, D., Staveley, T. A., and Gullström, M. 2018. Habitat connectivity of fish in temperate shallow-water seascapes. Frontiers in Marine Science, 4: 1–12.

Phan, H., Chén, O. Y., Pham, L., Koch, P., De Vos, M., McLoughlin, I., and Mertins, A. 2019. Spatio-temporal attention pooling for audio scene classification. *In* Proceedings of INTERSPEECH.

Picheral, M., Colin, S., and Irisson, J.-O. 2017. EcoTaxa'a tool for the taxonomic classification of images. http://ecotaxa.obs-vlfr.fr/. (Last accessed on October 2021).

Piczak, K. J. 2015. Environmental sound classification with convolutional neural networks. *In* Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, pp. 1–6.

Pine, W. E., Pollock, K. H., Hightower, J. E., Kwak, T. J., and Rice, J. A. 2003. A review of tagging methods for estimating fish population size and components of mortality. Fisheries, 28: 10–23.

Politikos, D. V., Fakiris, E., Davvetas, A., Klampanos, I. A., and Papatheodorou, G. 2021. Automatic detection of seafloor marine lit-ter using towed camera images and deep learning. Marine Pollution Bulletin, 164: 111974.

Rasmussen, J. H., and Širović, A. 2021. Automatic detection and classification of baleen whale social calls using convolutional neural networks. The Journal of the Acoustical Society of America, 149: 3635–3644.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. 2016. You only look once: unified, real-time object detection. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788.

Ristic, B., Arulampalam, S., and Gordon, N. 2004. Beyond the Kalman Filter: Particle Filters for Tracking Applications, Artech House Radar Library, Artech House. https://books.google.no/books?id=cjFDngEACAAJ.

Russell, S., and Norvig, P. 2002. Artificial intelligence: a modern approach. Englewood Cliffs, N.J: Prentice Hall.

Saba, G. K., Burd, A. B., Dunne, J. P., Hernández-León, S., Martin, A. H., Rose, K. A., Salisbury, J. *et al.* 2021. Toward a better understanding of fish-based contribution to ocean carbon flux. Limnology and Oceanography, 66: 1639–1664.

Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., and Battaglia, P. 2020. Learning to simulate complex physics with graph networks. *In* Proceedings of the International Conference on Machine Learning. PMLR, pp. 8459–8468.

Schiettekatte, N. M., Barneche, D. R., Villéger, S., Allgeier, J. E., Burkepile, D. E., Brandl, S. J., Casey, J. M. *et al.* 2020. Nutrient limitation, bioenergetics and stoichiometry: a new model to predict elemental fluxes mediated by fish. Functional Ecology, 34: 1857–1869.

Schmid, M. S., Daprano, D., Jacobson, K. M., Sullivan, C., Briseño-Avena, C., Luo, J. Y., and Cowen, R. K. 2021. 'A Convolutional Neural Network based high- throughput image classification pipeline – code and documentation to process plankton underwater imagery using local HPC infrastructure and NSF's XSEDE'. National Science Foundation, National Aeronautics and Space Administration, Belmont Forum, Extreme Science and Engineering Discovery Environment. doi: 10.5281/zenodo.4641158.

Schmidhuber, J. 2015. Deep learning in neural networks: an overview. Neural Networks, 61: 85–117.

Schneider, S., Taylor, G. W., Linquist, S., and Kremer, S. C. 2019. Past, present and future approaches using computer vision for animal re-identification from camera trap data. Methods in Ecology and Evolution, 10: 1151–1155.

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., and Carvalho, S. 2019. Chimpanzee face recognition from videos in the wild using deep learning. Science Advances, 5: 1–10.

Schroeder, S.-M., Kiko, R., and Koch, R. 2020. MorphoCluster: efficient annotation of plankton images by clustering. Sensors, 20: 3060.

Sharma, J., Granmo, O.-C., and Goodwin, M. 2020. Environment sound classification using multiple feature channels and attention based deep convolutional neural network. *In* Proceedings of the INTERSPEECH. 1186–1190. doi: 10.21437/Interspeech.2020-1303.

Shin, Y., Choi, J. H., and Choi, H. S. 2021. Deep learning based fish object detection and tracking for smart aqua farm. The Journal of the Korea Contents Association, 21: 552–560.

Shiu, Y., Palmer, K., Roch, M. A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T. *et al.* 2020. Deep neural networks for automated detection of marine mammal species. Scientific Reports, 10: 1–12.

Sun, Y., Xue, B., Zhang, M., and Yen, G. G. 2019. Completely automated cnn architecture design based on blocks. IEEE Transactions on Neural Networks and Learning Systems, 31: 1242–1254.

Suryanarayana, I., Braibanti, A., Rao, R. S., Ramam, V. A., Sudarsan, D., and Rao, G. N. 2008. Neural networks in fisheries research. Fisheries Research, 92: 115–139.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D. *et al.* 2015. Going deeper with convolutions. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Szesciorka, A. R., Ballance, L. T., Širović, A., Rice, A., Ohman, M. D., Hildebrand, J. A., and Franks, P. J. 2020. Timing is everything:

drivers of interannual variability in blue whale migration. Scientific Reports, 10: 1–9.

Tam, J. C., Link, J. S., Rossberg, A. G., Rogers, S. I., Levin, P. S., Rochet, M.-J., Bundy, A. *et al.* 2017. Towards ecosystem-based management: identifying operational food-web indicators for marine ecosystems. ICES Journal of Marine Science, 74: 2040–2052.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. 2020. Efficient transformers: a survey. a rXiv:2009.06732.

Tessler, C., Givony, S., Zahavy, T., Mankowitz, D., and Mannor, S. 2017. A deep hierarchical approach to lifelong learning in minecraft. *In* Proceedings of the AAAI Conference on Artificial Intelligence, 31.

Uusitalo, L., Fernandes, J. A., Bachiller, E., Tasala, S., and Lehtiniemi, M. 2016. Semi-automated classification method addressing marine strategy framework directive (MSFD) zooplankton indicators. Ecological Indicators, 71: 398–405.

Vadziutsina, M., and Riera, R. 2020. Review of fish trap fisheries from tropical and subtropical reefs: main features, threats and management solutions. Fisheries Research, 223: 105432.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008.

Waddell, E. E., Rasmussen, J. H., and Širović, A. 2021. Applying artificial intelligence methods to detect and classify fish calls from the Northern Gulf of Mexico. Journal of Marine Science and Engineering, 9: 1128.

Wang, N., Yu, J., Yang, B., Zheng, H., and Zheng, B. 2020. Vision-based in situ monitoring of plankton size spectra via a convolutional neural network. IEEE Journal of Oceanic Engineering, 45: 511–520.

Wang, Z. A., Moustahfid, H., Mueller, A. V., Michel, A. P., Mowlem, M., Glazer, B. T., Mooney, T. A. *et al.* 2019. Advancing observation of ocean biogeochemistry, biology, and ecosystems with cost-effective in situ sensing technologies. Frontiers in Marine Science, 6: 519.

Weinstein, B. G. 2018. A computer vision for animal ecology. Journal of Animal Ecology, 87: 533–545.

Whitt, C., Pearlman, J., Polagye, B., Caimi, F., Muller-Karger, F., Copping, A., Spence, H. *et al.* 2020. Future vision for autonomous ocean observations. Frontiers in Marine Science, 7: 697.

Wiggins, S. M., and Hildebrand, J. A. 2007. High-frequency acoustic recording package (HARP) for broad-band, long-term marine mammal monitoring. *In* Proceedings od the 2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies. IEEE, pp.551–557.

Wiggins, S. M., and Hildebrand, J. A. 2016. Long-term monitoring of cetaceans using autonomous acoustic recording packages. *In* Listening in the Ocean, Springer, pp. 35–59.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N. *et al.* 2019. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 6: 160018.

Wilson, R. J., Speirs, D. C., Sabatino, A., and Heath, M. R. 2018. A synthetic map of the North-West European shelf sedimentary environment for applications in marine science. Earth System Science Data, 10: 109–130.

Winn, H. E. 1964. The biological significance offish sounds. *In*Ed. by Tavolga, W. N., Marine bio-acoustics, pp. 213–231. Pergamon Press, Oxford.

Yang, X., Zhang, S., Liu, J., Gao, Q., Dong, S., and Zhou, C. 2021. Deep learning for smart fish farming: applications, opportunities and challenges. Reviews in Aquaculture, 13: 66–90.

Yu, X., Wei, Y., Zhu, M., and Zhou, Z. 2016. Automated classification of zooplankton for a towed imaging system. *In* Proceedings of the OCEANS 2016. IEEE, Shanghai.

Yu, Y., Zhao, J., Gong, Q., Huang, C., Zheng, G., and Ma, J. 2021. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. Remote Sensing, 13: 3555.

Zhang, Z., Xu, S., Zhang, S., Qiao, T., and Cao, S. 2020. Attention based convolutional recurrent neural network for environmental sound classification. Neurocomputing. 453: 896–903.

Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., and Brewer, A. 2020. Beluga whale acoustic signal classification using deep learning neural network models. The Journal of the Acoustical Society of America, 147: 1834–1841.

Handling Editor: David Demer