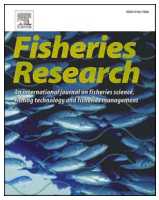
Fisheries Research 254 (2022) 106411

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



Evaluation of echosounder data preparation strategies for modern machine learning models

Alba Ordoñez [a,*], Ingrid Utseth [a], Olav Brautaset [a], Rolf Korneliussen [b], Nils Olav Handegard [b]

[a] Department of statistical analysis, machine learning and image analysis, Norwegian Computing Center, Gaustadalléen 23A, 0373 Oslo, Norway
[b] Department of ecosystem acoustics, Institute of Marine Research, Nordnesgaten 50, 5005 Bergen, Norway

ARTICLE INFO

Handled by A.E. Punt

Keywords:
Acoustic target classification
Multi-frequency echograms
Deep learning
Semantic segmentation
U-Net
Resampling
Sandeel

ABSTRACT

Fish stock assessment and management requires accurate estimates of fish abundance, which are typically derived from echosounder observations using acoustic target classification (ATC). Skilled operators are regularly assisted in classifying acoustic targets by software and there has been an increasing interest toward using machine learning to create improved tools. Recent studies have applied deep learning approaches to acoustic data, however, algorithm data-preparation strategies (influencing model output) are presently poorly understood and standardization is needed to enable collaborative research and management. For example, a common preprocessing technique is to resample backscatter data coming from echosounder measurements from the original resolution to a coarser resolution in the horizontal (time) and vertical (range) directions. Using data values derived from the volume backscattering coefficient obtained during the Norwegian sandeel survey, we investigate which resampling resolutions are suitable for ATC using a convolutional neural network trained to classify single values of backscatter data. This process is known as pixel-level semantic segmentation. Our results indicate that it is possible to downsample the data if important information related to acoustic characteristics is not smoothed out. We also show that the classification performance is improved when providing the network with contextual information relating to range. These findings will provide input to fisheries acoustic data standards and contribute to the on-going development of automated ATC methods.

1. Introduction

Acoustic surveys are a key component of pelagic fish stock assessment and management. Data are typically collected using hull-mounted, downward-looking echosounders, which provide echoes from organisms in the water-column (MacLennan, 1990). Data variables obtained from echosounders can relate to the observed backscatter, such as the volume backscattering coefficient, s_v, which can be attributed to the target fish species. The s_v data are integrated over a depth range (echo integration) and in accordance with the principle of linearity, is assumed to scale linearly to fish abundance (Foote, 1983). Other backscatter data variables such as the mean target strength can then be used to estimate fish abundance. Typically, abundance estimates, which in most cases are used as a relative index, are along with catch statistics used to parameterize fish stock assessment models, but some stocks rely solely on acoustic-based abundance estimates, e.g. capelin and sandeel.

A prerequisite to echo integration is acoustic target classification (ATC), see Horne (2000) and Korneliussen et al. (2018) for a review. This process allocates observed high-resolution backscatter data (displayed as an echogram) to species (e.g. herring, sandeel, etc.) or species groups (e.g. swimbladdered fish, zooplankton, etc.) by defining acoustic categories. This allocation has been coined the "holy grail of fisheries acoustics" (MacLennan and Holliday, 1996). The most common approach is to perform this manually, but automated methods that use a software desktop application exist. Typically the observed backscatter for the target species is visualized by frequency (Korneliussen and Ona, 2001; Kloser et al., 2002) and used to aid manual classification. Common software applications for this purpose are Echoview (Hobart, Tasmania) and LSSS (Marec, Norway; Korneliussen et al., 2016). Such manual approaches involve expert knowledge and lead to concerns about consistency among operators, hence the need for automated methods which are both reproducible and consistent.

Traditional approaches to the automation of target classification are reliant on the characterization of distinct echogram features such as schools and layers. In both cases the relative frequency response (Korneliussen and Ona, 2002) is an important feature. Information can be

* Correspondence to: Norwegian Computing Center, Gaustadalléen 23A, 0373 Oslo, Norway.
E-mail address: albao@nr.no (A. Ordoñez).

https://doi.org/10.1016/j.fishres.2022.106411
Received 15 February 2022; Received in revised form 21 June 2022; Accepted 22 June 2022
Available online 28 June 2022

extracted from the high resolution s_v data (e.g. morphometric features of fish schools; Haralabous and Georgakarakos, 1996), by averaging over many pings and samples to characterize the frequency response of distinct features (regions/layers) prior to classification (Simmonds et al., 1996), or by combining relative frequency response and morphology (Korneliussen et al., 2009). For methods based on the frequency response, one common pre-processing step is to downsample the s_v data from the original resolution to a coarser resolution to ensure that the acoustic beams from different transducers cover the same volume (Korneliussen et al., 2008). The challenge with these preprocessing approaches is that fine-scale information that could be used to aid target classification (Rose and Leggett, 1988) may be lost due to the averaging process.

Modern machine learning methods, e.g. deep convolutional neural networks (CNNs), can be used to extract important patterns/features from high resolution multi-dimensional data and perform automatic *end-to-end* learning (LeCun et al., 2015). Over the last few years, such methods have also emerged in marine sciences (Malde et al., 2019). For ATC, approaches were developed to classify the entire echogram (Hirama et al., 2017), perform pixel-level semantic segmentation (Brautaset et al., 2020) and instance segmentation (Marques et al., 2021) in backscatter data, as well as patch-level semantic segmentation based on semi-supervised learning (Choi et al., 2021).

However, no insights have yet been provided into how the averaging pre-processing of backscatter data (e.g. Korneliussen et al., 2008) can affect the outcome of ATC based on deep learning. Networks that can classify echograms (by treating them as images) may be able to learn any averaging filter or any other filter that improves classification, provided that the depth of the network is sufficient and the size of the filter covers enough samples.

In some instances, auxiliary information that is not part of the echogram (or image itself) can be used to further improve the performance of a network. Among the different available alternatives, concatenating extracted features from auxiliary data to the last layers of the original network provides a simple and efficient approach, which has demonstrated successful results in different domains including image classification (Tang et al., 2015), detections of defects for the textile industry (Calderisi et al., 2019) and plankton classification (Ellen et al., 2019). A network receiving as input groups of samples (i.e. patches) from an echogram (e.g. Brautaset et al., 2020) is not provided with the range information related to the different patches. To account for the range-dependent effect, i.e. that more individual targets may be covered as the beam widens by range, the network could be trained using the echogram together with information about the vertical range (depth) of the sample. It could be of interest to evaluate whether this auxiliary information could provide useful contextual information to the network about the increased amount of averaging in range that exists due to the acquisition.

The main objective of this paper is to test if resampling resolution affects ATC performed with neural networks trained for pixel-level semantic segmentation (a pixel referring here to a sample in range for a single ping). A secondary objective is to develop a method to include auxiliary information (e.g., related to the amount of averaging in range due to the acquisition) in such a CNN designed for acoustic data. Based on backscatter data from the Norwegian sandeel survey, the following three data preparation strategies were tested:

1. Follow the approach used in Brautaset et al. (2020) to train a CNN to classify the observed backscatter using data derived from high resolution s_v (baseline model).
2. Same as (1) but follow Korneliussen et al. (2008) and resample the s_v data according to their recommendation prior to training the CNN (resampled models).
3. Same as (1) but add auxiliary information regarding range in the CNN in two different ways (auxiliary models). In the first approach, we simply add the auxiliary data in the first layer of the CNN and in the second approach we integrate the auxiliary data to a later stage of the CNN as described in Tang et al. (2015).

A framework to compare the performance of the three data preparation strategies was developed and non-parametric statistical tests for pairs of models were carried out. The goal is to provide input to fisheries acoustics data standards for processing data used in modern machine learning models.

2. Materials and methods

2.1. Echosounder data

Our test case data were derived from high resolution s_v values, displayed as echograms, provided by the Institute of Marine Research, Norway, from the sandeel survey in the North Sea during spring. The acoustic data have been collected since 2005 (Johnsen et al., 2017) and our study was based on surveys acquired between 2007 and 2018, corresponding to the same data used in Brautaset et al. (2020) for classifying single values of backscatter data into acoustic categories using semantic segmentation. More specifically, the data measurements were done using a Simrad EK60 operating at 18, 38, 120, and 200 kHz. The ping rates varied within and across years (Fig. A. 1) with mean values ranging from 1.25 ping s^{-1} to 3.67 ping s^{-1}. The test case data had the highest resolution available from the echosounder output and was obtained by averaging over many (EK60-internal) samples to obtain an approximately constant vertical resolution of 18 cm across the data.[1]

Each sample of the echogram was manually labelled by the same operator across the entire period using the LSSS system (Marec, Norway; Korneliussen et al., 2016). For the pixel-level semantic segmentation process, the echograms were treated as images composed of pixels. As done in Brautaset et al. (2020), the multi-frequency s_v values that were organised into three-dimensional matrices (here denoted as *tensors*) in range, ping and frequency were logarithmically transformed and multiplied by 10 to obtain volume backscattering strength values, S_v, that were fed into the neural network models after thresholding (see Brautaset et al., 2020 for further details).

In addition to the S_v data, we considered the sample range (from the surface) as auxiliary information for the neural networks.

2.2. Neural network models

To evaluate whether ATC based on pixel-level semantic segmentation was affected by various data preparation strategies such as resampling and the addition of auxiliary data, we set up the following three different cases: 1.) Baseline, 2.) Resampled and 3.) Auxiliary.

2.2.1. Baseline

We used the same pixel-level semantic segmentation architecture proposed by Brautaset et al. (2020), which was based on the end-to-end convolutional U-Net model (Ronneberger et al., 2015). This architecture took input patches of size 256×256×4 samples; each patch consisting of 256 samples in the time dimension x 256 samples in the range dimension x 4 frequency channels (18, 38, 120, and 200 kHz) obtained from S_v values. The output of the network was acoustic category, which had the three different factor levels: “sandeel”, “background” and “other”.

2.2.2. Resampled

This case followed the same setting as the baseline, but prior to the conversion to S_v values, the s_v data were interpolated such that they shared either a common horizontal (time between pings) resolution or a

[1] The EK60-internal samples of 2–3 cm (depending on frequency) are not available to users, but are averaged internally in EK60 to give samples 1/4 length of the pulse, which in this case was 18 cm.

common vertical (range) resolution or a common horizontal-vertical grid. Each data point in the new grid was computed as a weighted mean of data points from the original grid, the weighting being the area of the intersection for the new grid-box with each of the original grid-boxes. The resampling process used the *AreaWeighted* function in the Iris Python package (Met Office, 2013). This resampling algorithm is first order conservative, preserving the area-weighted total backscatter energy in the data. Thus, echo integration of the backscatter data will still produce fish stock abundance estimates comparable to those based on the original, full-resolution data. The labels, in the form of two-dimensional matrices with the same number of samples as the original data, were resampled using nearest neighbour interpolation.

We considered two new horizonal grids with constant ping rates of 1 ping s^{-1} and 0.1 ping s^{-1}. The new vertical grid had a resolution of 50 cm. By selecting lower horizontal- and vertical resolutions for the new grids we downsampled the s_v data as suggested in Korneliussen et al. (2008). Resampling horizontally only (2 grids), vertically only (1 grid) and both horizontally and vertically (2 grids) led to five new resampled S_v datasets as illustrated in Fig. 1.

By resampling the data to a constant horizontal distance-grid, i.e. based on ship-log or GPS, which is equivalent to constant distance assuming constant survey speed, the objective was to make multi-ping features more comparable, since the ping rate of the original data varied across years. Note that in addition to varying number of pings per distance due to varying survey speed, a varying ping-rate is also common due to varying bottom-depth. Downsampling to a lower vertical resolution of 50 cm across years allowed for input patches containing the entire water column in almost all instances during training and prediction (as 99.7% of the pings had a seabed shallower than 256×50 cm = 128 m), without an increase in memory usage.

2.2.3. Auxiliary

This case followed the same setting as the baseline with the addition of the range auxiliary variable to the CNN using two different strategies. First, we simply concatenated the auxiliary data with the multi-frequency S_v data to obtain input patches of size 256×256×5 that were fed to the U-Net model (Fig. 2a). For that, we turned the auxiliary data into a tensor of shape 256×256×1 pixels, where each pixel was assigned its corresponding range (i.e. vertical number of samples from the surface). This case was denoted as *early auxiliary*. Second, following Tang et al. (2015) we integrated extracted features from the auxiliary data tensor at a higher semantic level of the network trained with S_v data (Fig. 2b) and referred to this as *late auxiliary*. To extract meaningful features from the auxiliary variable, we utilized a small, separate convolutional neural network consisting of two layers of 32 convolutions with 1×1 kernels (linear projections) followed by a rectified linear activation function (ReLU). The number of kernels were chosen for efficiency of execution and to be able to keep the same batch size as in Brautaset et al. (2020) during training. This avoided extending the training time compared to the baseline case and demanding more memory than available for common graphics processing units (GPUs), which were used during training. After using the last convolutional layer characterized by one kernel of size 1×1, the extracted features from the separate network were concatenated to the features extracted from the S_v data via the U-Net layers. The concatenation occurred at the level before the final convolutional and *softmax* layers of the U-Net model, which computed the vector of probabilities for the augmented output allowing the final classification of the pixels.

When integrating auxiliary data in the U-Net architecture, the learning process of the network was more stable when the auxiliary variable values were between 0 and 2 or slightly over (we used patches of data selected above the seabed and in 95% of the cases the seabed had a vertical extension of 500 samples). Therefore, we divided the auxiliary

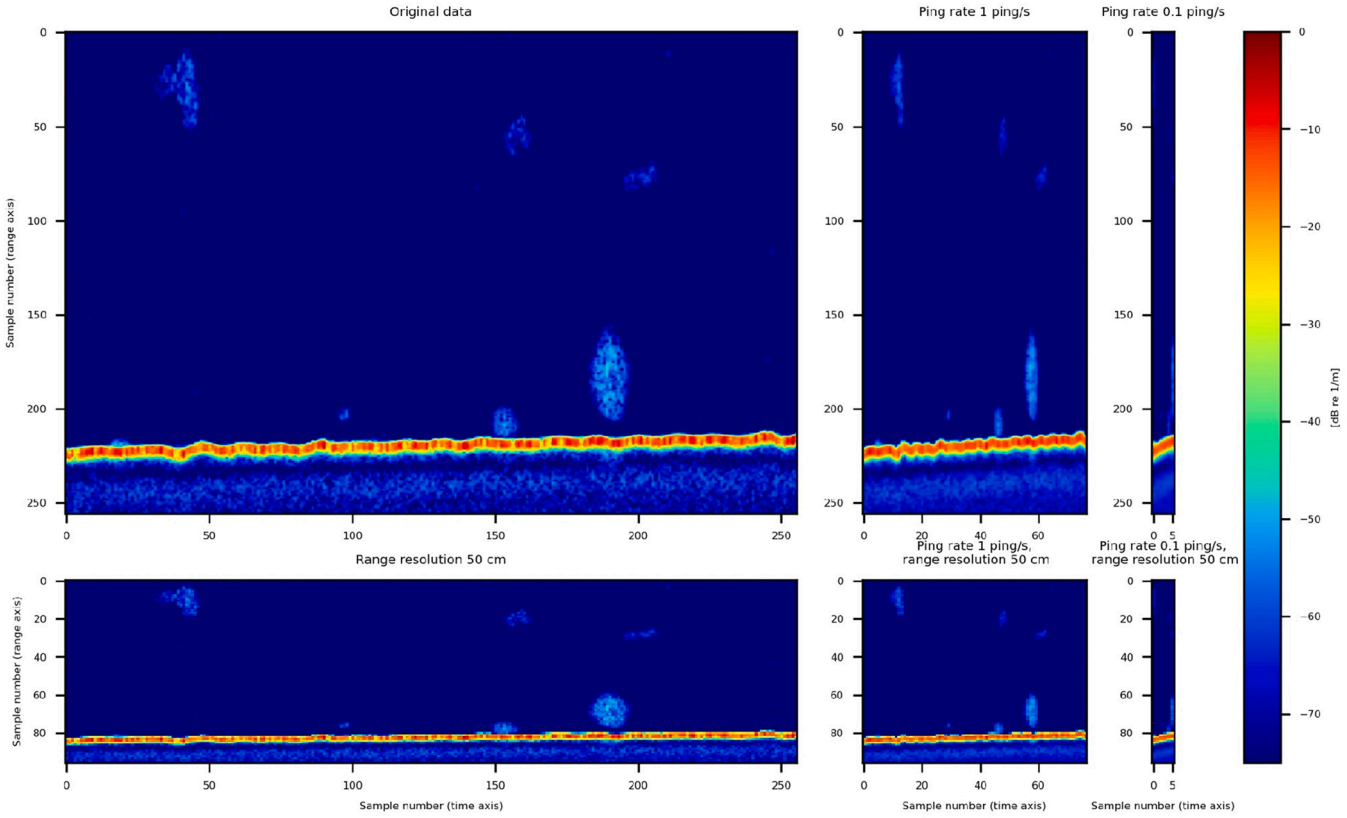


Fig. 1. A sample of S_v data (200 kHz channel) before resampling (upper left corner) and after resampling with 5 different grids. The original data patch had 256×256 samples, a mean ping rate of 3.3 ping s^{-1} and a range (vertical) resolution of 18 cm.

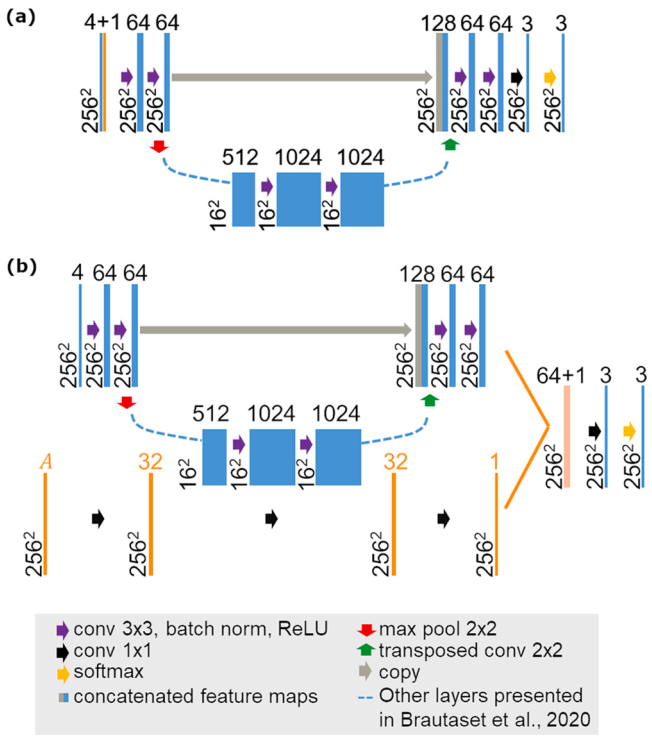


Fig. 2. Network architectures incorporating auxiliary range data. (a) Early auxiliary model: the original U-Net network presented in Brautaset et al. (2020) was modified such that the auxiliary data was simply concatenated to the multi-frequency S_v data leading to an input data tensor of size 256×256 x (4 + 1). (b) Late auxiliary model: the original U-Net network was modified such that extracted features from the auxiliary data tensor were concatenated to the penultimate layer of the original model. The tensor containing auxiliary data was sent into a separate simple network (bottom orange branch) composed of 1×1 kernel convolutions, each followed by a ReLU activation function. The extracted features of size 256×256×1 were concatenated to the features extracted from the multi-frequency S_v data (256×256×64). The concatenation occurred at a high semantic level of the original U-Net model.

variables by 256, which was the vertical extent of the input tensor used during training.

2.2.4. Training and evaluation procedures

To evaluate which data pre-processing strategy would work better for ATC based on pixel-level semantic segmentation, we set up a framework to better assess whether there were potential improvements in the baseline model performance when using resampled and auxiliary models. We fixed a common random seed and used the same training framework for the different experiments. As no information on this was available in the model from Brautaset et al. (2020), we could not directly use that model as our baseline. As we had to re-train the baseline, we took this opportunity to slightly modify some parameters compared to the original paper.

Consistent with the original paper, we trained the model using a batch size of 16 and optimized with stochastic gradient descent, but we used a different initial learning rate of 0.005 (reducing this value every 1000 iterations by a factor of 2) and momentum value 0.95. We used early stopping, imposed a limit of training iterations of 10 000 and weighted the chosen cross-entropy loss with class weights ("background"=10, "sandeel"= 300, and "other"=250 selected after experiencing with different value parameters on randomly chosen echograms using the baseline model). To deal with the imbalance present in the dataset (the number of background pixels greatly outnumbering those of the two additional classes) we used the same sampling strategy as described in Brautaset et al. (2020). Each trained model was exposed to 16,000 randomly chosen samples of width and height equal to 256 selected across the same training years as chosen in Brautaset et al. (2020) (i.e. data acquired between 2011 and 2016 and excluding 2012). We ran four trials for each of the models, where a different random seed was chosen each time (but was common to all the different strategies for a given run). This allowed us to check whether the results were obtained by chance (consistency of the model). Finally, the predictions were averaged before the softmax layer of the U-Net architecture. The experiments were implemented using the PyTorch framework on GTX 1080 Ti GPU devices.

To assess the ability of the different models to detect the sandeel schools, we proceeded as in Brautaset et al. (2020): we computed precision and recall curves per year considering "sandeel" as the positive class and "background" together with the "other" class as negative. For the prediction implementation, we followed Brautaset et al. (2020) and applied the trained models to small overlapping image patches for obtaining the classification results at the pixel level. This included a post-processing step where all pixels more than 10 pixels below the seabed were classified as "background" (to correct for potential misclassified pixels from the initial prediction, e.g. from false echoes). Note that for the baseline and auxiliary cases, the predictions were kept at the same resolution as the original multi-frequency S_v data, whereas for the resampled cases, we used the same resolution as the interpolated datasets used for training.

The precision and recall curves were obtained by considering all the pixels from the echograms within a single year. As we used data acquired between 2007 and 2018, we obtained 11 curves in total (data from the year 2012 did not have the 120 kHz frequency channel and was therefore ignored). These curves were used to extract the maximized F1-scores (derived from the harmonic mean of both measures) and allowed us to compare the performance for pairs of models, considering baseline, resampled and auxiliary scenarios.

3. Results

From the obtained F1-score distribution by model considering all the data from different surveys (Fig. 3), we observed a general improvement in F1-score compared to the baseline when training on vertically

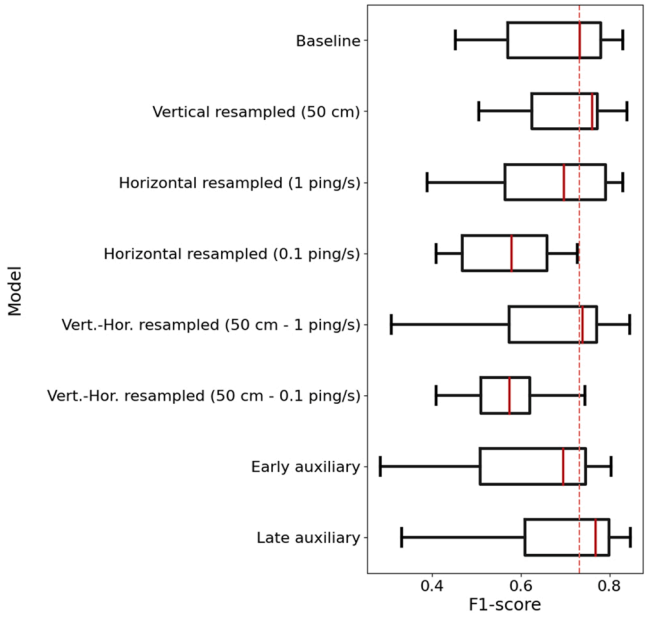


Fig. 3. Boxplots displaying the F1-score distribution by model considering years 2007–2018. Whiskers are positioned according to Tukey's original definition of boxplots (Tukey, 1977). The red dashed line highlights the median baseline F1-score.

resampled data. This was also overall confirmed across the years when examining the precision and recall curves (Fig. A. 2). When using the late auxiliary model incorporating range data, the model seemed to further improve (Fig. 3). Resampling to 1 ping s^{-1} led to a lower median F1-score with respect to the baseline. The models trained on data involving a resampling to a common horizontal grid of 0.1 ping s^{-1} yielded clear deteriorations of the F1-score results. The same occurred when using the early auxiliary model. For those 3 models (horizontal resampled 0.1 ping s^{-1}, vert.-hor. resampled 50 cm – 0.1 ping s^{-1} and early auxiliary), the precision and recall curves (Fig. A. 2) consistently showed lower recall values than the baseline for all the surveys, except the 2007 one.

One sided Wilcoxon non-parametric tests were performed for pairs of models, to test whether the differences in performance seen in the boxplots (Fig. 3) were statistically significant. More specifically, we computed the differences in F1-scores between reference models (baseline and late auxiliary) and the rest of the models across the different years using the information presented in Table A.1. This allowed us to test whether these differences were significantly larger/smaller than zero. We accounted for test multiplicity by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995), which is the expected proportion of false positives among all rejected null

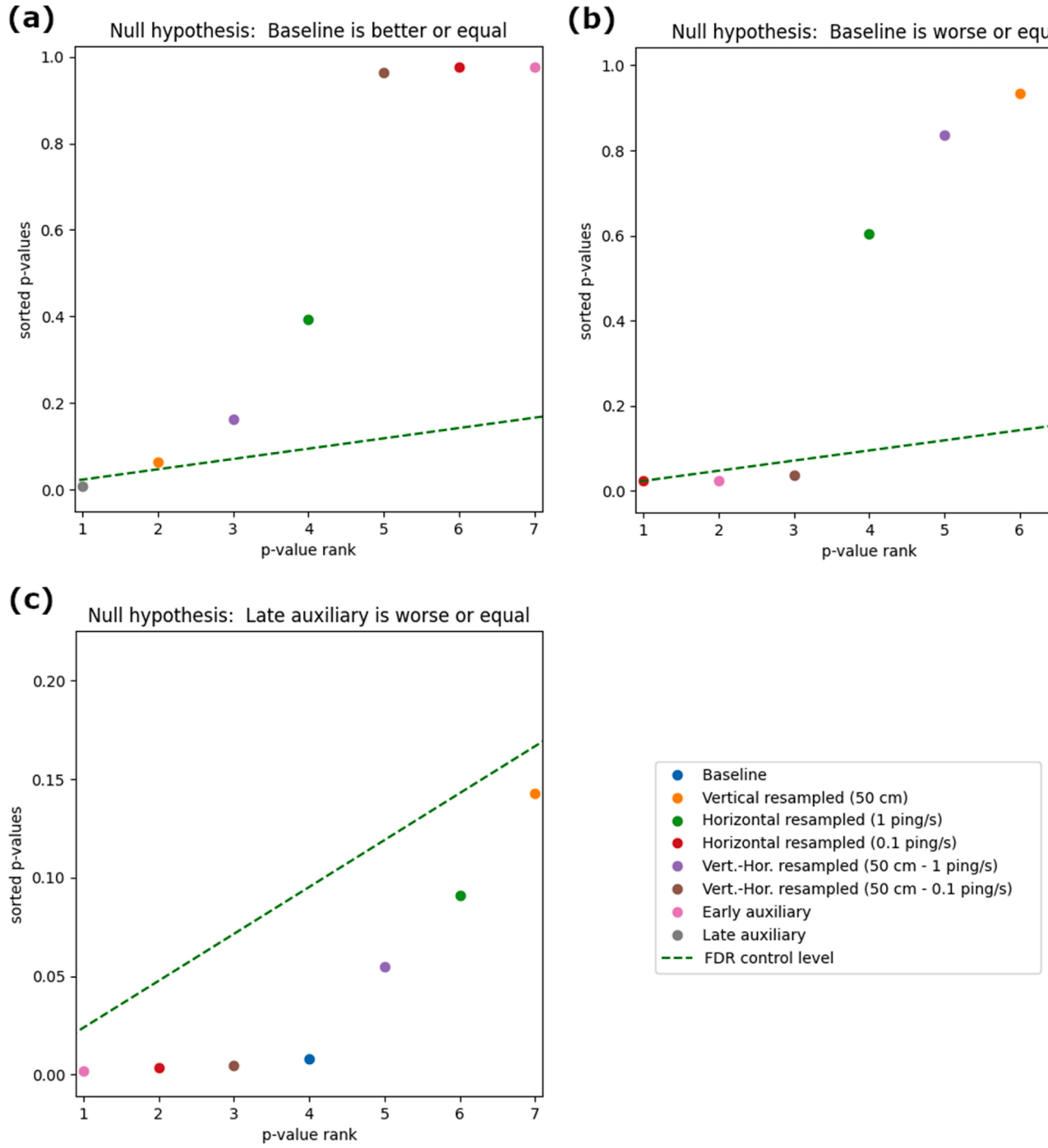


Fig. 4. Illustration of different statistical tests carried out using the FDR criterion by choosing a control level of 1/6 (dashed line). The p-values are related to the differences in performance between pairs of models and the p-value ranks correspond to the comparisons of pairs of models. For the points below the control level, we have statistical significance that we can reject the null hypothesis.

hypotheses. The FDR control level was fixed at $\alpha_{FDR} = 1/6$ such that if, for example, 6 out of the 7 null hypotheses in our comparison were rejected, we would expect that at most one of these rejections was erroneous. The results are illustrated in Fig. 4 and summarized in Table 1. The improvement in F1-score relative to the baseline when downsampling the data vertically (50 cm) was not statistically significant (Fig. 4a) although the associated p-value was very close to the control level. The results were significantly degraded when using a horizontal resampling of 0.1 ping s^{-1} or the early auxiliary model (Fig. 4b). Resampling to 1 ping s^{-1} did not significantly degrade the baseline results (Fig. 4b). When using the late auxiliary model, the improvement in performance relative to the baseline and the other models was also significant (Fig. 4a, Fig. 4c).

The examination of a few qualitative examples in terms of predictions highlighted that the resampling strategies and the addition of auxiliary data generally improved the baseline results, in the presence of dense plankton layers near the surface (Fig. 5a, Fig. A. 3a-b). However, resampling strategies led to distorted morphology of the predicted sandeel schools compared to the baseline (Fig. 5b, Fig. A. 3c-d), sometimes missing small schools near the seabed (Fig. 5b), which seemed to be less the case with the late auxiliary model.

4. Discussion

There are a range of different echosounder settings that may affect ATC, like power, ping repetition intervals etc., but the objective of this paper was to test whether we could cast the backscatter data into a tensor for ATC based on deep learning and evaluate the effect of various data preparation strategies on the performance of a model trained to detect sandeel schools. This included testing different resampling resolutions and the addition of auxiliary data related to the amount of averaging in range due to the acquisition. The echo-sounder system was calibrated, and the ping-rate was as high as practically possible during data collection. In addition, to the best of our knowledge there was no crosstalk between channels. The transmit pulse-duration was set to 1 ms to maintain time-series. Our results showed that keeping the original ping resolution while resampling the data vertically to fit a three-dimensional tensor was a viable approach for pre-processing the backscatter data. In that case, a general improvement in F1-score compared to the baseline was observed and the increase in performance was close to being statistically significant. Not pre-processing the backscatter data but providing the network with contextual information about range, statistically improved performance when the auxiliary data was injected at a higher semantic level of the deep learning network. These findings could have implications when it comes to providing a best practice recommendation on how to use standard deep learning frameworks developed for acoustic backscatter images.

Earlier papers have shown that averaging the frequency response horizontally and vertically stabilizes the performance (Simmonds et al., 1996; Korneliussen et al., 2008). Since the convolutions span several pings and range bins, the CNN can, in theory, learn any filter (and combination of filters), which could minimize the need for manual data pre-processing. If averaging is important for performance, it is likely that the first layers in the CNN have learned how to average the data prior to classification. The need for averaging is likely less important for CNNs than for the more traditional approaches and has the added benefit that fine scale information may be retained for classification purposes. This allows for utilizing fine scale structures, e.g. as those utilized by Rose and Leggett (1988), in addition to the coarser (averaged) scales. This is the advantage of end-to-end learning where the fine scale patterns in the non-prepocessed backscatter data may be utilized simultaneously with the coarser scales (since the filters extracting features from the input data can learn both scales).

For a given network size, larger water volumes can be input into the CNN by averaging the data horizontally at the expense of reducing information related to finer-scale features. Our results indicated that important information related to acoustic characteristics of the detected sandeel schools was present at finer scales, since degradation in performance became statistically significant compared to the baseline when resampling the data at coarser scales. One explanation for the decrease in model performance with coarser (horizontal) grid size is that averaging at scales approaching the size of typical sandeel schools (median height of 21 pixels and width of 6 pixels in the original S_v data) smoothed out part of the school signal containing ping-level fine scale structures that were important. Indeed, the median width of sandeel schools considering all years from 2007 to 2018 was reduced to 1 pixel when horizontally resampling to 0.1 ping s^{-1} or when resampling both vertically to 50 cm and horizontally to 0.1 ping s^{-1}. For these data that were resampled, a larger part of the signal may have been misclassified as background.

Vertical downsampling did not affect the results to the same degree and the increase in performance was close to being statistically significant. The median height of sandeel schools decreased from 21 pixels (original data) to 7 pixels when vertically downsampling to 50 cm, suggesting that the vertical scale of most features was coarser than the range interval (approximately 18 cm). This is somewhat contradictory to the findings of Rose and Leggett (1988) that reported that the finer structures in range were important for classification, although they considered other species than sandeel. Since the survey area was shallow (maximum seabed depth in 99.7% of the cases did not exceed 128 m), the improved performance may be attributed to the grid covering a larger extension of the water column, allowing the network to obtain information about absolute location in range (the scale structures ping to ping remained preserved).

Similarly to sandeel, Atlantic mackerel have such a characteristic acoustic signature that a network may not underperform despite a resampling in the range and/or time directions. For other species like herring or capelin, however, the higher variation existing in the s_v data (Korneliussen et al., 2016) may be important to classify the fish schools. Thus, finding the most appropriate resampling resolution to distinguish these species from others using deep learning could be an important task in the future.

The original range (vertical) resolution of the echosounder data was determined by the pulse duration. When a similar pulse duration is used across frequencies, the data can be added to a standardized tensor. In our data set, a few years had settings that deviated from the standard settings and performing a vertical resampling to the same grid allowed us to add all data to the same standardized tensor. For pulse-compressed data, e.g. as implemented in the Simrad EK80, the effective pulse-duration varies with bandwidth, which means that the number of samples per ping will be different between different transducers. Since our results

Table 1
Summary of the non-parametric statistical tests when comparing the baseline scenario to the resampling and auxiliary scenarios.

Scenario	Statistically better than baseline	Statistically worse than baseline	No worse, no better than baseline
Vertical resampling (50 cm)			X
Horizontal resampling (1 ping s^{-1})			X
Horizontal resampling (0.1 ping s^{-1})		X	
Vert.-Hor. resampling (50 cm – 1 ping s^{-1})			X
Vert.-Hor. resampling (50 cm – 0.1 ping s^{-1})		X	
Early auxiliary		X	
Late auxiliary	X		

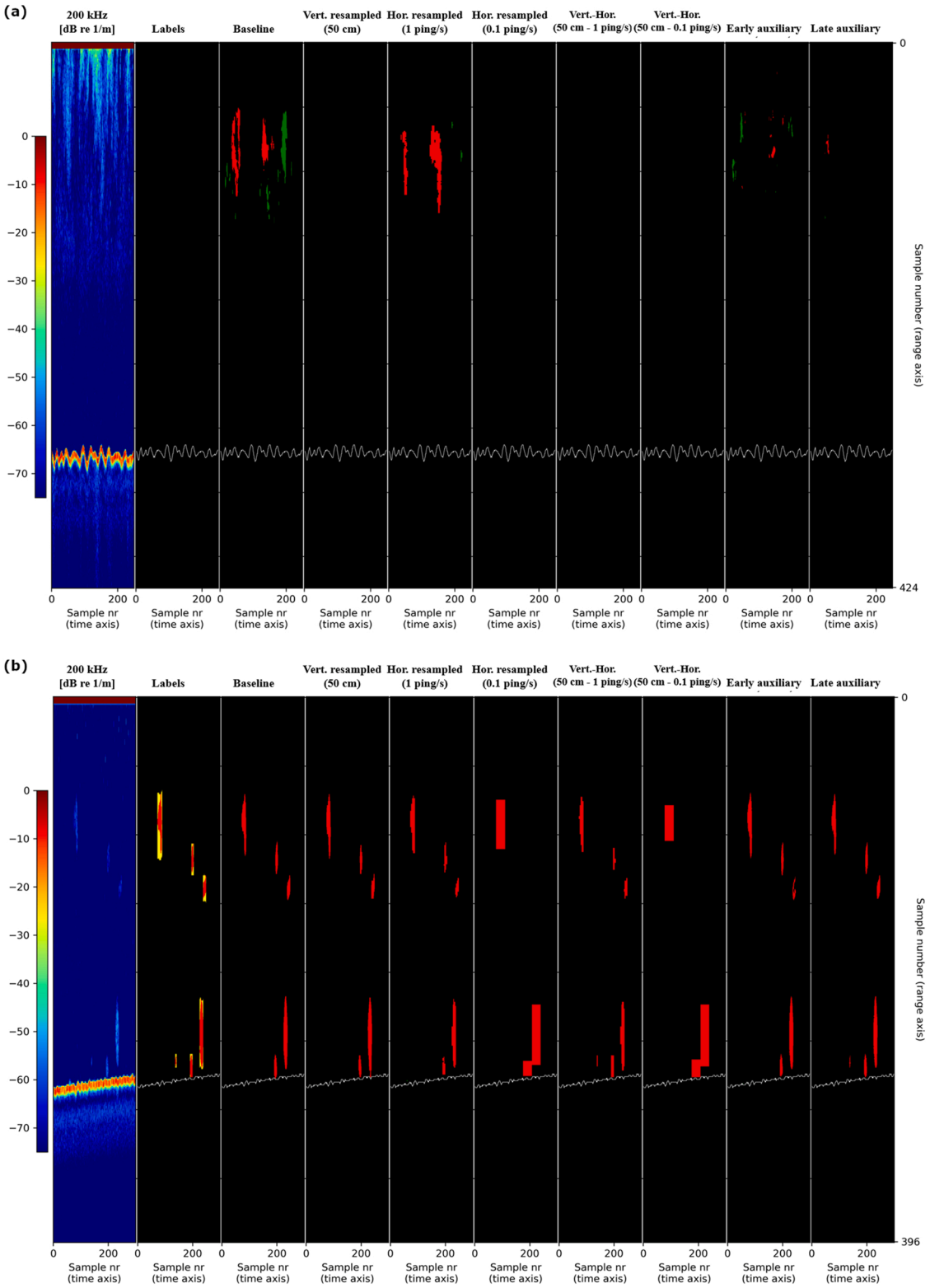


Fig. 5. Qualitative examples showing the 200 kHz S_v data, followed by the annotated labels and the predictions obtained using different models. Note that for visualization purposes the predictions of the resampled models were upsampled to the original resolution. (a) Area with dense plankton layers at the surface and no sandeel schools, (b) area with several sandeel schools, with a small sandeel school near the seabed. The "background" class, the "sandeel" class and the "other" class are displayed in black, red and green respectively. As the annotations were somewhat coarse and most annotated school regions contained some background pixels as well, edge pixels of annotated school regions with low intensity were ignored during training (marked here as yellow).

showed that the network performance was not affected by vertical resampling, casting the data into a common tensor for easy interfacing with standard deep learning frameworks seems reasonable. This will also allow backward compatibility between Simrad EK60 and EK80 data. An alternative to casting the data into a common tensor would be to handle each channel as a separate input layer, enabling different resolutions to be used. In this case, the time for each ping would have to be included as auxiliary information so that the network could be designed or learn how to align the pings.

The strategy to use the grid configuration as an input and include range as an auxiliary variable at a higher semantic level gave the best improvement in performance over the baseline. Using this auxiliary variable provided the CNN with additional information about range effects (i.e. data at shallower ranges look less smoothed than at deeper ranges), allowing to adjust the initial U-Net predictions. In this context, it could be interesting to evaluate whether the proposed approach for including auxiliary data would help improving detection of other species than sandeel. Another benefit of handling the grid through auxiliary variables was that the input data was kept as close as possible to the observed backscatter data, such that fine-scale information could be utilized together with the coarser scales. Note that for the sake of keeping training and evaluation times as close as possible to the original U-Net, we chose a rather simple architecture to incorporate auxiliary information at a higher semantic level of the network. Simply concatenating auxiliary and S_v data and use these as input into the network statistically degraded the baseline performance. This was probably because the combined data were too different from each other to be handled with a U-Net network and alternative architectures may be taken into account going forward.

5. Conclusions

The main conclusions from the work are that: (i) it is possible to grid the backscatter data to fit deep learning algorithms designed for acoustic classification, (ii) contextual information in range can be important for network performance. We have tested this on the sandeel survey data and further work would be to test this on other survey time series where deep learning models are being developed. The benefit of casting the data into a tensor is that a wide range of image-based methods can be directly applied on acoustic data, accelerating the development of fully automated pipelines and bringing the holy grail of ATC within reach.

CRediT authorship contribution statement

Alba Ordoñez: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, visualization. **Ingrid Utseth**: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, visualization. **Olav Brautaset**: Software, Validation, Writing – review & editing. **Rolf Korneliussen**: Conceptualization, Data curation, Writing – review & editing, Supervision. **Nils Olav Handegard:** Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is a part of the CRIMAC and COGMAR projects, funded by the Norwegian Research Council (grants 309512 and 270966) that we would like to thank.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2022.106411.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57, 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X.

Brautaset, O., Waldeland, A.U., Johnsen, E., Malde, K., Eikvil, L., Salberg, A.-B., Handegard, N.O., 2020. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. ICES J. Mar. Sci. 77, 1391–1400.

Calderisi, M., Galatolo, G., Ceppa, I., Motta, T., Vergentini, F. 2019. Improve Image Classification Tasks Using Simple Convolutional Architectures with Processed Metadata Injection. *In* 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 223–230.

Choi, C., Kampffmeyer, M., Handegard, N.O., Salberg, A.B., Brautaset, O., Eikvil, L., Jenssen, R., 2021. Semi-supervised target classification in multi-frequency echosounder data. ICES J. Mar. Sci.

Ellen, J.S., Graff, C.A., Ohman, M.D., 2019. Improving plankton image classification using context metadata. Limnol. Oceanogr. Methods 17, 439–461.

Foote, K.G., 1983. Linearity of fisheries acoustics, with addition theorems. J. Acoust. Soc. Am. 73, 1932–1940.

Haralabous, J., Georgakarakos, S., 1996. Artificial neural networks as a tool for species identification of fish schools. ICES J. Mar. Sci. J. du Cons. 53, 173–180.

Hirama, Y., Yokoyama, S., Yamashita, T., Kawamura, H., Suzuki, K., Wada, M. 2017. Discriminating fish species by an Echo sounder in a set-net using a CNN. *In* 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), pp. 112–115.

Horne, J.K., 2000. Acoustic approaches to remote species identification: a review. Fish. Oceanogr. 9, 356–371.

Johnsen, E., Rieucau, G., Ona, E., Skaret, G., 2017. Collective structures anchor massive schools of lesser sandeel to the seabed, increasing vulnerability to fishery. Mar. Ecol. Prog. Ser. 573, 229–236.

Kloser, R.J., Ryan, T., Sakov, P., Williams, A., Koslow, J.A., 2002. Species identification in deep water using multiple acoustic frequencies. Can. J. Fish. Aquat. Sci. 59, 1065–1077.

Korneliussen, R. Berger, L., Campanlla, F., Chu, D., Demer, D., De Robertis, A., et al. 2018. Acoustic Target Classification. ICES Cooperative Research Report No. 344.

Korneliussen, R.J., Ona, E. 2001. Some applications of multiple frequency echo sounder data. In Proceedings of the Scandinavian Symposium on Physical Acoustics 2001, pp 78–81.

Korneliussen, R.J., Ona, E., 2002. An operational system for processing and visualizing multi-frequency acoustic data. ICES J. Mar. Sci. 159, 293–313.

Korneliussen, R.J., Diner, N., Ona, E., Berger, L., Fernandes, P.G., 2008. Proposals for the collection of multifrequency acoustic data. ICES J. Mar. Sci. 65, 982–994.

Korneliussen, R.J., Heggelund, Y., Eliassen, I.K., Johansen, G.O., 2009. Acoustic species identification of schooling fish. ICES J. Mar. Sci. 66, 1111–1118.

Korneliussen, R.J., Heggelund, Y., Macaulay, G.J., Patel, D., Johnsen, E., Eliassen, I.K., 2016. Acoustic identification of marine species using a feature library. Methods Oceanogr. 17, 187–205.

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

MacLennan, D.N., 1990. Acoustical measurement of fish abundance. J. Acoust. Soc. Am. 87, 1–15 (Acoustical Society of America).

MacLennan, D.N., Holliday, D.V., 1996. Fisheries and plankton acoustics: past, present, and future. ICES J. Mar. Sci. 53, 513–516.

Malde, K., Handegard, N.O., Eikvil, L., Salberg, A.-B., 2019. Machine intelligence and the data-driven future of marine science. ICES J. Mar. Sci.

Marques, T.P., Cote, M., Rezvanifar, A., Albu, A.B., Ersahin, K., Mudge, T., Gauthier, S. 2021. Instance Segmentation-Based Identification of Pelagic Species in Acoustic Backscatter Data. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4373–4382.

Met Office, 2013. Iris: A Python package for analysing and visualising meteorological and oceanographic data sets. Exeter, Devon.

Ronneberger, O., Fischer, P., Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical image computing and computer-assisted intervention, pp. 234–241.

Rose, G.A., Leggett, W.C., 1988. Hydroacoustic signal classification of fish schools by species. Canadian Journal of Fisheries and Aquatic Sciences. NRC Research Press, Ottawa, Canada. Accessed 9 June 2021. ⟨https://cdnsciencepub.com/doi/abs/10.1139/f88-073⟩.

Simmonds, J.E., Armstrong, F., Copland, P.J., 1996. Species identification using wideband backscatter with neural network and discriminant analysis. ICES J. Mar. Sci. 53, 189–195.

Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., Bourdev, L. 2015. Improving Image Classification with Location Context. *In* Proceedings of the IEEE international conference on computer vision 2015, pp 1008–1016.

Tukey, J.W., 1977. Exploratory data analysis. Reading, Mass. Addison-Wesley Pub. Co.