




Contribution to the Themed Section: 'Applications of machine learning and artificial intelligence in marine science'

Original Article

Acoustic classification in multifrequency echosounder data using deep convolutional neural networks

Olav Brautaset¹, Anders Ueland Waldeland¹, Espen Johnsen², Ketil Malde², Line Eikvil¹, Arnt-Børre Salberg¹, and Nils Olav Handegard ^{2*}

¹Norwegian Computing Center, P.O. Box 114 Blindern, Oslo 0314, Norway

²Institute of Marine Research, Nordnesgaten 50, Bergen 5005, Norway

*Corresponding author: tel: + 47 95854057; e-mail: nilsolav@hi.no.

Brautaset, O., Waldeland, A. U., Johnsen, E., Malde, K., Eikvil, L., Salberg, A.-B., and Handegard, N. O. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. – ICES Journal of Marine Science, 77: 1391–1400.

Received 5 July 2019; revised 5 November 2019; accepted 14 November 2019; advance access publication 21 January 2020.

Acoustic target classification is the process of assigning observed acoustic backscattering intensity to an acoustic category. A deep learning strategy for acoustic target classification using a convolutional network is developed, consisting of an encoder and a decoder, which allow the network to use pixel information and more abstract features. The network can learn features directly from data, and the learned feature space may include both frequency response and school morphology. We tested the method on multifrequency data collected between 2007 and 2018 during the Norwegian sandeel survey. The network was able to distinguish between sandeel schools, schools of other species, and background pixels (including seabed) in new survey data with an F1 score of 0.87 when tested against manually labelled schools. The network separated schools of sandeel and schools of other species with an F1 score of 0.94. A traditional school classification algorithm obtained substantially lower F1 scores (0.77 and 0.82) when tested against the manually labelled schools. To train the network, it was necessary to develop sampling and preprocessing strategies to account for unbalanced classes, inaccurate annotations, and biases in the training data. This is a step towards a method to be applied across a range of acoustic trawl surveys.

Keywords: acoustic classification, big data, deep learning, machine learning, sandeel

Introduction

Acoustic trawl surveys (Simmonds and MacLennan, 2005) are commonly used in fisheries assessments to provide data that support advice on total allowable catches for a wide range of fish stocks. Echosounders are instruments that produce soundwaves and record the intensity of backscattered soundwaves produced by targets in the water column. Echosounder observations are calibrated (Foote *et al.*, 1987) and can be integrated over specific depth ranges to calculate the area-backscattering coefficient (the average backscattering intensity per metre square; MacLennan *et al.*, 2002). The area-backscattering coefficient is linearly related to fish abundance (Foote, 1983) for a given representative target strength (Ona, 2003), and under the assumption that the backscattering intensity can be correctly assigned to a species or a

group of species. The categorization of species into groups is aided by trawl samples of the fish, which are used to estimate the age and length distributions of fish populations. This method is typically used for pelagic or semi-pelagic species, such as walleye pollock (Karp and Walters, 1994), herring, blue whiting (Gastauer *et al.*, 2016), capelin (Gjosæter *et al.*, 2015), and sandeel (Johnsen *et al.*, 2009).

The process of assigning values of acoustic backscattering intensity to an acoustic category or group is typically a manual operation. An operator, based on information discerned from trawl catches, multifrequency echosounder observations, and any other auxiliary information, assigns values of acoustic backscattering intensity to an acoustic category, which can represent a species or a group of species. The process is typically time-

consuming and often incurs operator-based biases (Simmonds and MacLennan, 2005). To reduce bias and increase efficiency, several features ascertained from the acoustic observations have been used to aid, automate, or partially automate the process (Korneliussen, 2018). In addition to trawl sampling, features such as the location and position, environmental variables, and acoustically derived morphometric and energy features may also have discriminatory power (e.g. Horne, 2000; Reid, 2000). The main feature used in species classification is the relative frequency response, i.e. the fraction of backscattering intensity observed at one frequency relative to a reference frequency, typically 38 kHz (Kloser et al., 2002; Korneliussen and Ona, 2003). Based on these features, different methods have been used to classify values of backscattering intensity, including Bayesian methods (Korneliussen et al., 2016), semi-supervised methods (Woillez et al., 2012), and machine learning methods including random forest (Proud et al., 2020; Fallon et al., 2016) and artificial neural networks (Haralabous and Georgakarakos, 1996).

The current methods require that the feature space used for the classification is predefined, e.g. averaging the relative frequency response over a suitable number of pixels or defining the most efficient morphometric features, but this step is not trivial, i.e. how much should we smooth and what are the best morphological shapes? Defining the feature space for broadband fisheries echosounders (Mukai and Amakasu, 2016), where small-scale features in the frequency response may have large discriminatory power, may be even more challenging. A method that combines the feature extraction with the classification is preferable.

In recent years, deep convolutional neural networks (CNNs) have emerged as the leading modelling tools for image classification, segmentation, and semantic mapping both generally (Hariharan et al., 2015; Long et al., 2015) and also within marine science (Malde et al., 2020). CNNs do not require features to be designed in advance as they can learn the appropriate features from “raw” data, like images, and they have been shown to be superior in solving problems in computer vision and image analysis (Russakovsky et al., 2015). A CNN consists of a sequence of operations, referred to as *layers*, applied to the input image. The output from one layer is thus the input to the subsequent layer. Each layer typically consists of a number of separate convolutions with small *filter kernels*, followed by some non-linear function, and may also be combined with other operations. Each filter kernel consists of a number of coefficients, and using gradient-based optimization, these filter coefficients are tuned to minimize the classification error on annotated training data, referred to as *training* (Rumelhart et al., 1986). During training, the first layers will typically learn to recognize edges, lines, and corners, and the later layers can represent more abstract features. With this approach, the network can use the raw data directly as opposed to the traditional approach where the features must be predefined. Training a CNN requires large amounts of training data, i.e. ground truth data with corresponding annotations.

Image segmentation using CNNs can be carried out using several different approaches. One strategy is to train a classifier on small image patches and then either classify all pixels using a sliding window approach, or more efficiently, by converting the fully connected layers in the CNN to convolutional layers (Sermanet et al., 2013), thereby avoiding overlapping computations. Another approach is pixel-to-pixel semantic mapping using end-to-end learning (Long et al., 2015). It uses a fully convolutional network (FCN), consisting of an encoder and a decoder, where

the encoder maps the image to a low-resolution representation and the decoder provides a mapping from the low-resolution representation to the pixel-wise representation. An FCN has the advantage that the input size can vary since the convolutions “slide” over the data set, as opposed to networks that have fully connected layers requiring a fixed input size. A popular network architecture for semantic mapping is the U-Net (Ronneberger et al., 2015), characterized by skip connections between the corresponding encoder and decoder layers.

The objectives of this article are to (i) develop a deep learning strategy that is suitable for segmenting and classifying echosounder data collected during acoustic trawl surveys without prior feature extraction; (ii) demonstrate that the strategy developed in (i) works on a real test case, and (iii) provide perspectives, e.g. pros and cons, on the use of deep learning algorithms in the classification of acoustic observations into acoustic categories (e.g. species groups).

Material and methods

The sandeel survey

Data collected during the Norwegian North Sea Sandeel survey were used as test case for this study (ICES, 2016). The lesser sandeel (*Ammodytes marinus*), hereafter sandeel, is a small fish that does not have a swim bladder. For large parts of its life the sandeel hides by burrowing in sandy seabed, where the proportion of fine silt and clay particles is low (Macer, 1966; Wright et al., 2000). During the feeding season in spring, adults that have burrowed into the sandy substrate at night emerge at dawn (Winslade, 1974) to form large schools in the upper pelagic zone and predate on zooplankton (Freeman et al., 2004; Johnsen et al., 2017). The sandeel is a key species in the North Sea ecosystem, being a major prey species for several predators, including sea birds, seals, and larger fish (Furness, 2002), and is also a valuable target for commercial fishing.

The Institute of Marine Research, Norway, has been conducting acoustic trawl surveys for sandeel during April and May since 2005 in the sandeel areas of the north eastern part of the North Sea (Johnsen et al., 2017). The survey series (2005–2018) was conducted using the RV Johan Hjorth (2005–2008, 2010–2011), RV GO Sars (2009), FV Brennholm (2012), and FV Eros (2013–2018). All vessels were equipped with multifrequency Simrad EK60 echosounder systems operating transducers at 18, 38, 120, and 200 kHz, except for the FV Brennholm (2012) that was without a 120-kHz EK60 echosounder but collected 120 kHz using a Simrad ME70 sonar (Trenkel et al., 2008). In addition, the RV GO Sars and FV Eros (from 2014) were equipped with a 70- and 333-kHz echosounder. The echosounders were calibrated in accordance with standard procedures before each survey (Foote et al., 1987). During operation, the pulse duration and ping repetition frequency were set to 1.024 ms and 3–4 Hz, respectively, for all frequencies and vessel speed was kept at approximately ten knots. Echosounder observations were stored as values of volume backscattering coefficient (s_v ; average amount of backscattering intensity per cubic metre) by frequency (MacLennan et al., 2002). See Johnsen et al., 2009 for further details.

Data preprocessing

In some instances, the pulse duration (i.e. range resolution) and ping rate differed from the standard settings. To be consistent, the data were interpolated into a common time-range grid based

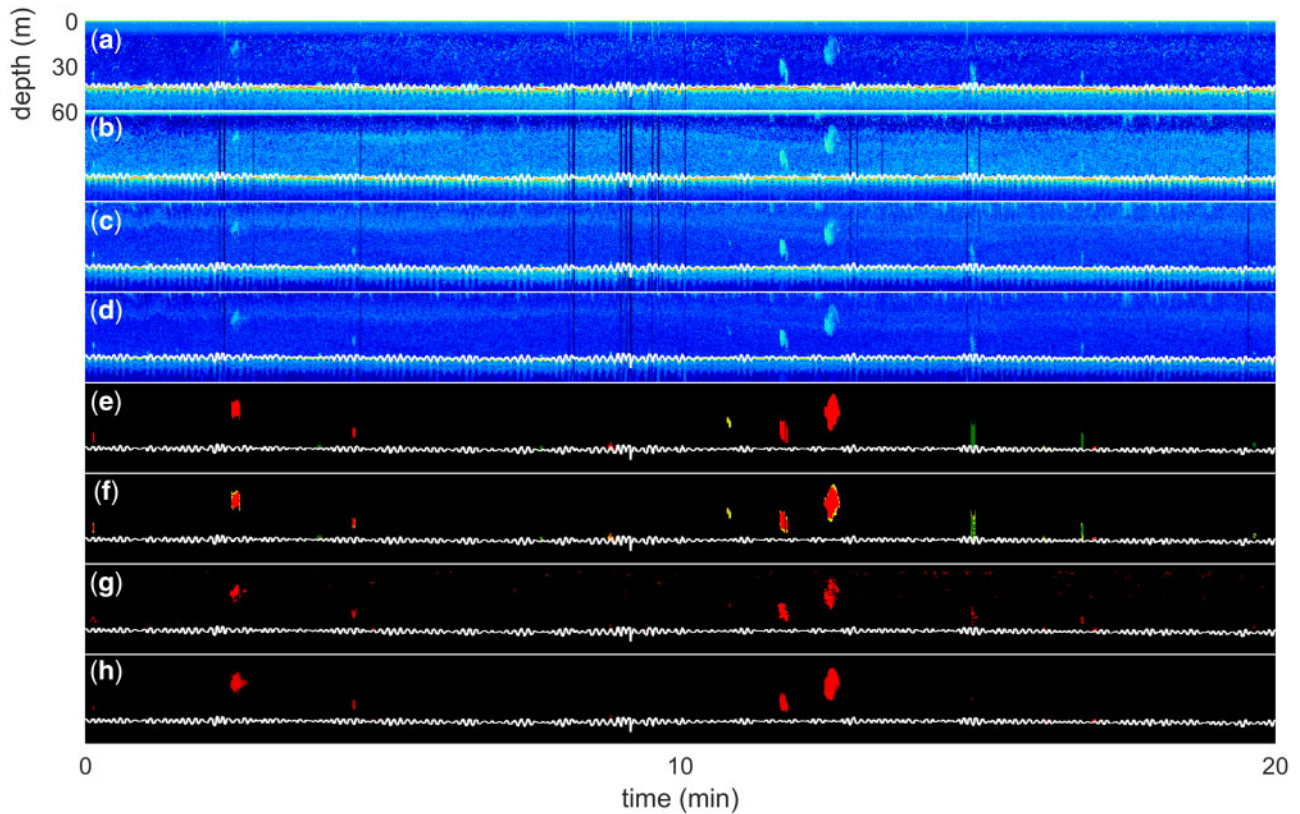


Figure 1. Echogram with four frequency channels (a–d, 18, 38, 120 and 200 kHz) and original (e) and modified (f) annotations, where black is the “background” class, red (grey in print) is the “sandeel” class, green (dark grey in print) is the “other” class, and yellow (light grey in print) is the “ignore” pseudo class, the predictions from the benchmark method (g), and the predictions from our method (h). Here, black and red (grey in print) are the background/other and sandeel classes, respectively. The seabed is shown as a white curve in all panels.

on the resolution of the 200-kHz data. The median ping rate was used to detect missing pings, and, when missing pings occurred, columns of zeros (mapped to -75 dB re 1 m^{-1} after log transformation) were inserted into the s_v data. If the range vector of the other frequencies was of a lower resolution, the data were interpolated onto the 200-kHz range vector. If the range vector had a higher resolution, the s_v values were averaged into bins defined by the 200-kHz range vector. This resulted in s_v values in a uniform time-range grid (Figure 1a–d), similar to pixels in a four-channel image, and we refer to these values as pixels hereafter. The seabed was approximately located as the depth with maximum increase in vertical gradient for each ping. This was used for balanced sampling (see below) and to avoid false predictions.

The survey series uses “sandeel”, “other”, “0-group sandeel”, and “possible sandeel” as acoustic categories, denoted “classes” hereafter, that were manually annotated by the same operator across all years. The annotations were interpolated into a pixel map corresponding to the echosounder data, and each pixel was allocated to one class. The acoustic classes “other” and “sandeel” have been used for all years, and the “sandeel” class is the only class used in official survey estimates. In addition, “possible sandeel” was introduced for schools where the frequency response was not consistent with sandeel but where the operator was in doubt and, for the 2016 survey, the “0-group sandeel” was introduced due to an extraordinary high density of juveniles. Each school varied from a few metres in length and height to >1 km in length extending across large parts of the water column (Johnsen

et al., 2017). The 200-kHz data were used as the primary frequency during annotation since it has the highest sandeel signal-to-noise ratio, and each school was annotated and classified by acoustic category using the Large Scale Survey System (LSS) postprocessing software (Korneliusson *et al.*, 2016). The manual annotations were mainly based on the frequency response of each school (see Johnsen *et al.*, 2009) and validated by trawl samples where applicable. The “0-group sandeel” and “possible sandeel” classes were added to an “ignore” pseudo class, and all other pixels not associated with a class were set to “background”. This resulted in pixel-based annotations with classes “sandeel”, “other”, “background”, and “ignore” (Figure 1e). Note that the bottom echo is included in the “background” class. Table 1 shows the total number of schools for each class.

The purpose of the annotations is to estimate sandeel abundance, which is calculated by summing up the 200-kHz backscattering intensity (Figures 1d and 2a) of the sandeels over a given region and dividing by their mean target strength. Heave measurements of the survey vessel were used to correct the echogram data and annotations. However, all figures are presented without heave corrections. The annotations were often coded as rectangular bounding boxes (when viewed with heave corrections; Figure 2b), and a portion of the bounding box would, consequently, include background pixels. This does not substantially affect the abundance estimate since adding low-value pixels does not substantially contribute to the total integrated backscattering intensity, but it may confuse a pixel-based classifier trying to

Table 1. Number of schools annotated as “sandeel”, “other”, and “ignore” per year in the final dataset.

Year	Sandeel schools	Other schools	Ignored schools
2007	453	605	0
2008	1 664	4 378	0
2009	699	2 755	30
2010	3 206	2 560	542
2011	623	1 685	177
2013	2 015	5 133	527
2014	1 121	6 113	549
2015	1 515	4 866	523
2016	829	4 423	2 130
2017	3 602	2 362	755
2018	4 678	1 917	255
Total	20 405	36 797	5 488

Table 2. Sampling strategy for drawing random $4 \times 256 \times 256$ crops for training.

Classes	Probability	Description
Background	1/26	Random crop from area without fish, above the seabed
Seabed	5/26	Random crop from area containing seabed
Sandeel	5/26	Random crop from area containing “sandeel” class
Other	5/26	Random crop from area containing “other” class
Seabed + sandeel	5/26	Random crop from area containing both seabed and “sandeel” classes
Seabed + other	5/26	Random crop from area containing both seabed and “other” classes

We divided regions of the echograms into these six classes and drew random samples from each class with the given probabilities.

predict the “background” class. To amend this problem, we modified the original annotations based on the s_v values. Any pixel annotated as “sandeel” or “other” with a corresponding 200-kHz s_v value outside the interval $[10^{-7} \text{ m}^{-1}, 10^{-4} \text{ m}^{-1}]$ was assigned to the “ignore” pseudo class (Figure 2). We set the threshold values based on a visual inspection of multiple echograms. We further smoothed the fish annotated pixel regions by applying binary morphological closing to the modified “sandeel” and “other” annotations, using a 7×7 disk-shaped structure element (Figure 2c).

Training data

For each survey, the acoustic data were comprised of a single continuous echogram for each frequency, but for training purposes, we divided each dataset into $4 \times 256 \times 256$ pixels crops, where 4 is the number of frequencies. We also applied a decibel transform to the s_v values and applied a hard threshold to values below $-75 \text{ dB re } 1 \text{ m}^{-1}$ and above $0 \text{ dB re } 1 \text{ m}^{-1}$. Each annotated echogram also possessed a heavy class imbalance; there were many more “background” pixels (99.8%) than “sandeel” (0.1%) and “other” pixels (0.1%). To expose the network to enough samples with fish schools when training, we first created an algorithm to get crops that were composed entirely of background pixels and similarly for crops that included “sandeel” and “other” pixels, respectively. We then balanced the dataset by applying an

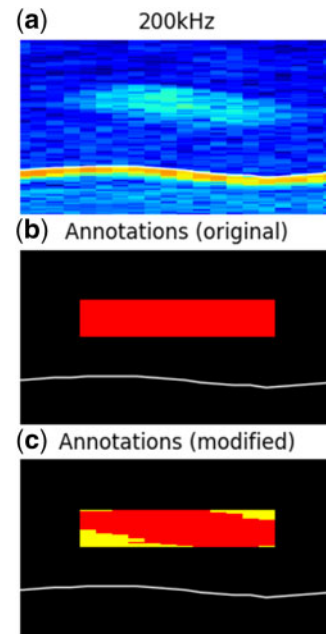


Figure 2. (a) Small patch from an echogram (200-kHz channel) with (b) original and (c) modified annotations. Modified annotations were obtained from original annotations using thresholds on the 200-kHz channel followed by morphological closing. The classes “background”, “sandeel”, and “ignore” are presented in black, red, and yellow (black, grey, and light grey in print), respectively. Axes are similar to Figure 1, where the vertical and horizontal axes represent depth and time, respectively.

equal sampling probability to crops containing seabed only, “sandeel”, “other”, seabed and “sandeel”, and seabed and “other” (see Table 2). All these crop types include the “background” class, but, in addition, we randomly sampled a smaller fraction of crops that had “background” pixels only (see Table 2). In addition, most of the sandeel schools resided close to the seabed and the balanced sampling during training mitigated the network from classifying all schools close to the seabed as sandeel, or worse, classifying the bottom itself as sandeel.

We partitioned the dataset into a training and validation dataset and a test dataset by different years, where 2011–2016 was used for training and validation and 2007–2010 combined with 2017–2018 was used for testing. From the training and validation set, we used 85% randomly drawn echograms for training and the remaining 15% for validation to select the best model. Among the test sets, the final year (2018) was unseen until the final evaluation.

Deep learning model and training

In this study, we built a classifier that was based on a slightly modified version of the U-Net architecture (Ronneberger *et al.*, 2015). The U-Net is a pixel-wise image segmentation network with a convolutional encoder–decoder architecture (Figure 3 and Supplementary Tables S1 and S2), originally developed for the segmentation of medical images. An encoder–decoder architecture can represent both pixel-wise and abstract features simultaneously. Our modified U-Net takes four frequency channels, 18, 38, 120, and 200 kHz, and a 256×256 range-time subset of the echogram as the input ($4 \times 256 \times 256$), and “encode” it to a

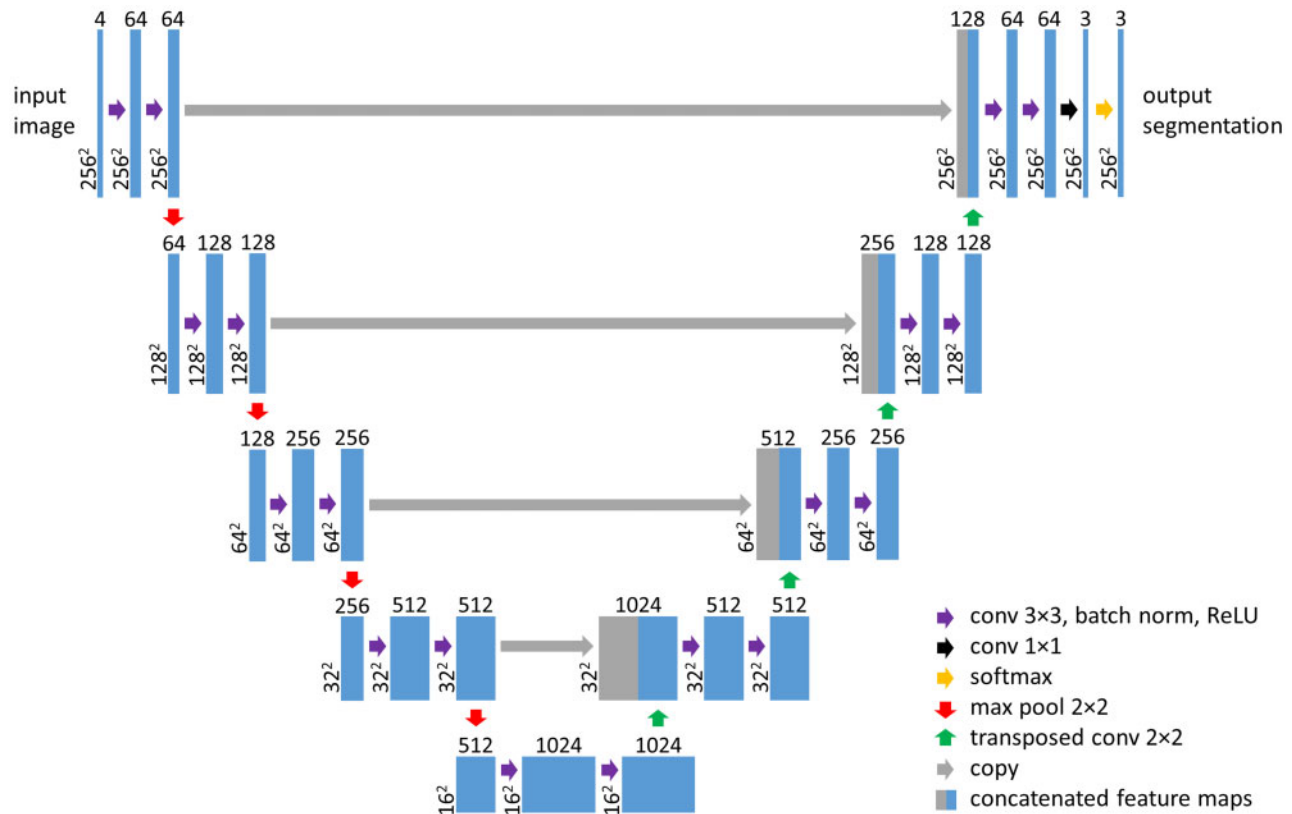


Figure 3. The network architecture, a slightly modified version of the original U-Net. The input is the $4 \times 256 \times 256$ crops, and the $3 \times 256 \times 256$ output is the softmax for each pixel by class (“sandeel”, “other”, and “background”).

16×16 “image” with 1024 abstract features ($16 \times 16 \times 1024$). The decoder then takes these features and generates (decode) an output for the classes “background”, “other”, and “sandeel” ($3 \times 256 \times 256$) for each of the input pixels. The architecture also copies the lower level features at each step when decoding, resulting in the decoder to both have access to low-level features (e.g. the frequency response in a small region) and more abstract features (e.g. like the overall shape). Finally, the output is passed through a “softmax” function where each of the three output classes is mapped to the interval $[0, 1]$ and add up to 1, like a probability for each class for each pixel. Contrary to the original implementation, we inserted a batch normalization layer (Loffe and Szegedy, 2015) between each convolutional layer and its subsequent activation function to reduce covariate shift, i.e. normalizing the distribution of outputs from each convolutional layer.

We trained the model over 5000 iterations using batches of 16 random $4 \times 256 \times 256$ crops. We used random uniform weight initialization and optimized with stochastic gradient descent with initial learning rate 0.01 and momentum 0.95. The learning rate controls how much the model parameters can change in each training iteration, while the momentum controls how much a training sample will influence the change of model parameters in the subsequent iterations. The learning rate was reduced by a factor of 0.5 every 1000 iterations. The model was evaluated on the validation set every 20th iteration. Due to the class imbalance, we used a weighted cross entropy loss with class weights (background = 1, sandeel = 30, and other = 25) to further adjust for imbalanced classes, giving less weight to each background pixel to

compensate for this class being more frequently observed. We randomly flipped the training crops about the vertical axis and added random multiplicative noise to a random 5% of the pixels. The hyperparameters were set by training the model multiple times, each time with a different combination of hyperparameters. We observed the impact on classification accuracies on the training and validation set for different combinations and fine-tuned the hyperparameters further based on the combination that gave the best initial results.

Since the network is based on convolutions only, the input image can be of any size during prediction and does not have to resemble the $4 \times 256 \times 256$ crops used for training. When using the network for prediction, we applied it to tiled segments (corresponding to the echosounder raw files) of the full survey echograms, including an overlap between segments of 40 pixels to avoid edge effects. As a postprocessing step, we also removed any predictions of fish more than ten pixels below the seabed.

Due to the heavy class imbalance, we used precision/recall curves rather than receiver operating characteristic curves to evaluate the performance. The network is considered a binary pixel classifier (positive/negative) by fixing a *threshold* value between 0 and 1, classifying a pixel as positive if the network output for the “sandeel” class is above this threshold value and negative otherwise. Using “sandeel” as the positive class, the precision is the proportion of *predicted* “sandeel” pixels that are correct, and the recall is the fraction of “sandeel” labels that are correctly predicted as “sandeel”. By predicting all pixels as “sandeel”, recall would be

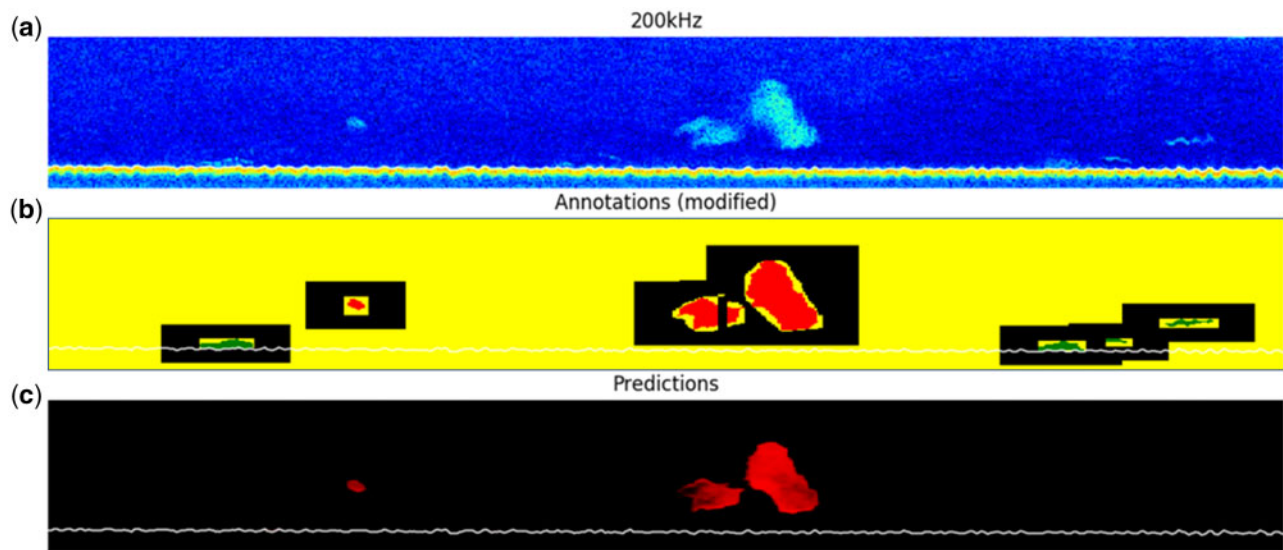


Figure 4. Illustration of evaluated pixels for computing precision/recall curves. (a) The 200-kHz echogram, (b) modified annotations where yellow pixels (light grey in print) are the ignore pseudo class, while in this example, “sandeel” (red, grey in print) is treated as positive and “other” (green, dark grey in print) and “background” (black) are regarded as negatives when calculating the precision/recall curves. (c) The predictions of the “sandeel” class where a high softmax is shown as bright red and a low softmax is shown as dark red (grayscale in print). Axes are similar to Figure 1, where the vertical and horizontal axes represent depth and time, respectively.

1, but precision would be low, and conversely, by correctly predicting one pixel as sandeel, precision would be 1, but recall would be low. Varying the threshold value results in different precision and recall values, where the recall may increase at the cost of lowering the precision. For a good classification, both precision and recall should be high and the F1 score at a given threshold value, defined as

$$F1 = \left(\frac{\text{precision}^{-1} + \text{recall}^{-1}}{2} \right)^{-1},$$

is typically used to test a methods performance. In our case, we only report the maximized F1 score, i.e. choosing the threshold value that gives the highest F1 score.

When evaluating the performance, we used two slightly different approaches when calculating the precision/recall curves for the background class. The first approach classifies echograms using all the echogram pixels, whereas the other approach evaluated echogram regions that were within 20 pixels of any original school annotation (c.f. Figure 4). The rationale behind using these two approaches was that we suspected that a proportion of schools was not classified during annotation, and therefore, comparing “sandeel” predictions to annotations for entire echograms may result in a high number of erroneous false positives. This would again yield poor precision/recall curves and not reflect the actual performance of the model.

When calculating the precision/recall curves, we used different combinations of classes as positives and negatives, i.e. “sandeel” as positive vs. “other” as negative to test the ability to separate species given a school is detected and “sandeel” vs. “other” and “background” to test the overall ability to detect sandeel schools, which is the purpose of the survey. Predictions of the “ignore” pseudo class were not considered when calculating the curves (c.f. Figure 4).

Evaluation

To test our approach against a traditional automated processing pipeline, we used the Sandeel case in Korneliussen *et al.* (2016) as a benchmark. This was implemented as a Korona processing pipeline in LSSS and consisted of a range of operations, including noise filtering (spike noise, spot noise) smoothing, bottom detection, thresholding, school detection, and categorization. We used the exact same setup and parameters as used by Korneliussen *et al.* (2016). The categorization was exported to a file and imported and treated similarly as the predictions from the U-Net algorithm, except that the threshold for accepting a pixel as sandeel was fixed, resulting in one point in the prediction recall plot as opposed to the curves from our method. The testing was only performed in the years that we used as test cases, i.e. our network had never seen the data where we compare the methods.

Results

We trained and validated the model using echograms derived from 2011 to 2016 survey data and tested the trained model using echograms derived from 2007 to 2010 and 2017 to 2018 survey data. Figure 1h shows an example of classification based on model predictions for a four-channel echogram. In this example, the trained network successfully separated the sandeel schools from other types of fish and the background class. Figure 1g shows the corresponding classification based on the benchmark method.

The network’s ability to discriminate “sandeel” (positive) vs. “background” and “other” (negative) is good (F1 score 0.87, Figure 5) when excluding background pixels that are at a distance of 20 pixels or more from the school annotations. In this case, a total number of 170 million pixels were evaluated (positives and negatives) and the annotations consisted of 90% “background”, 6% “sandeel”, and 4% “other” (Supplementary Table S3). This resulted in an overall F1 score of 0.87 for the overall test set across

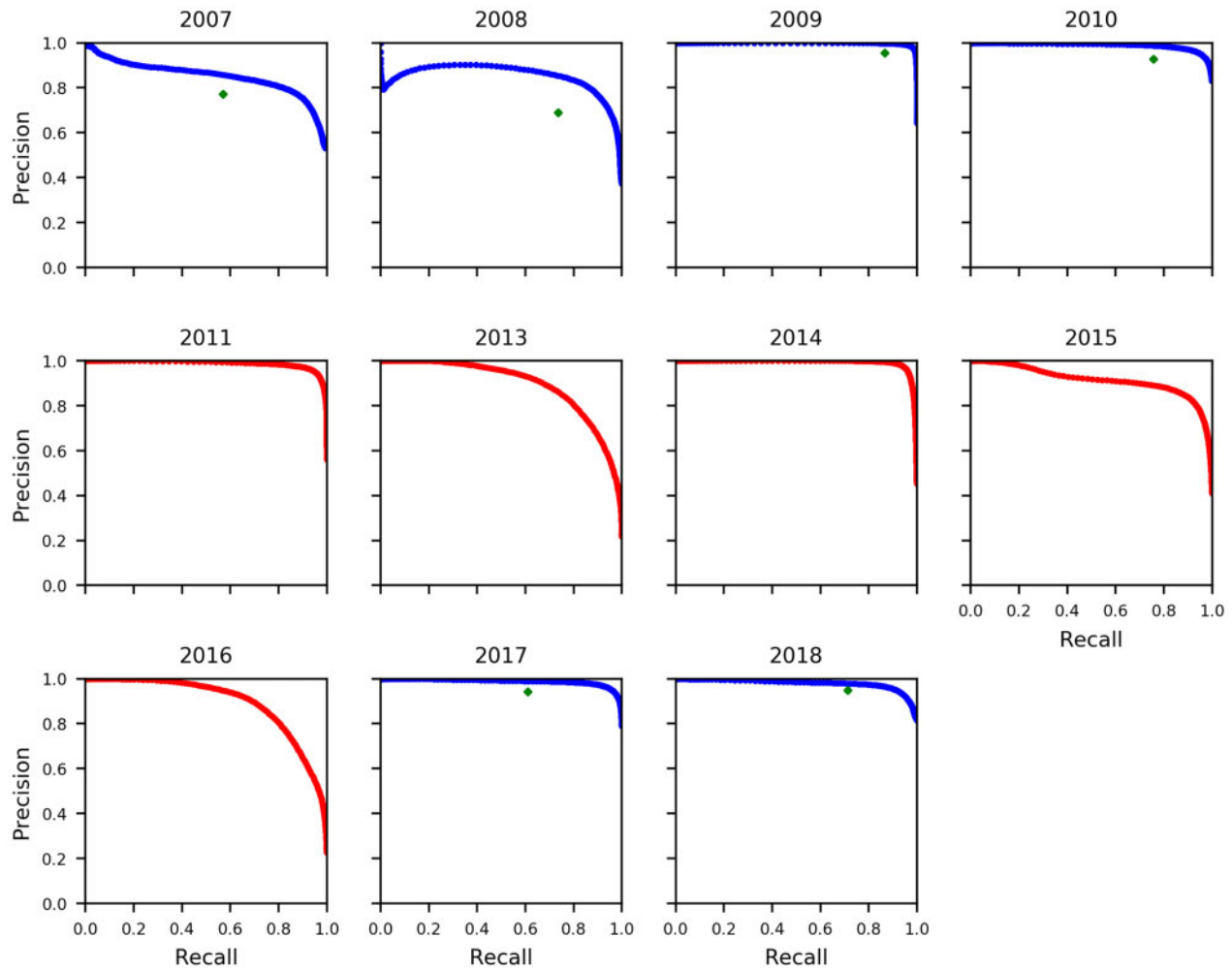


Figure 5. Precision/recall curves per year, where “sandeel” (positive) is compared to “other” and “background” (negatives) in a 20-pixel region extending beyond the original school annotations. The remaining pixels annotated as “background” or “ignore” were excluded. The red and blue curves (dark grey and light grey in print) denote the training data (years 2011–2016) and test data (years 2007–2010 and 2017–2018), respectively. Each diamond denotes the corresponding precision/recall value for the benchmark method (evaluated on test data years only).

years, with a corresponding threshold, precision, and recall of 0.80, 0.85 and 0.89, respectively. For the training and validation set, the years 2013, 2015, and 2016 did not perform as well when compared to the other years, and for the test set, the years 2007 and 2008 did not perform as well as the other years. The benchmark method achieved an overall F1 score of 0.77 for the overall test set across years, with a corresponding precision and recall of 0.80 and 0.74, respectively (Figure 5).

We also tested the network’s ability to discriminate between “sandeel” (positive) and “other” (negative) while excluding both “background” and the pseudo-class “ignore”, i.e. the ability to determine the species given that a school is detected. In this case, a total number of 18 million pixels were evaluated (positives and negatives) and the annotation consisted of 0% background, 57% “sandeel”, and 43% “other”. Our model’s separation of sandeel vs. other species obtained an overall F1 score of 0.94 for the test set. The corresponding threshold, precision, and recall were 0.50, 0.93, and 0.95, respectively. The test set results by year were also more consistent than the previous case (including background pixels), with the exception of 2007 and 2008, indicating that the network is well suited to differentiate between species

(Supplementary Figure S1). The benchmark method achieved an overall F1 score of 0.82 for the test set, with a corresponding precision and recall of 0.91 and 0.74, respectively (Supplementary Figure S1).

Our model did not perform as well when tested using entire echograms as input (Supplementary Figure S2). The performance on the test set for the years 2017 and 2018 was satisfactory (F1 score 0.61 and 0.78, respectively) but was substantially poorer for earlier years 2007–2010 (F1 score 0.11, 0.51, 0.78, and 0.68, respectively). The benchmark method achieved even lower F1 scores, both for the years 2017 and 2018 (0.32 and 0.62, respectively) and for the years 2007–2010 (0.03, 0.07, 0.42 and 0.50, respectively; Supplementary Figure S2). When looking into these specific results in more detail, we found two main reasons for the discrepancies, including missing annotations, incomplete annotations, and erroneous predictions close to the sea surface.

Missing annotations were found in several echograms, and an example of this is provided in Figure 6c, where the entire right-hand side of the echogram does not contain any annotations of fish. On closer inspection of the 200-kHz echogram (Figure 6a), clear fish marks were not annotated. In these circumstances,

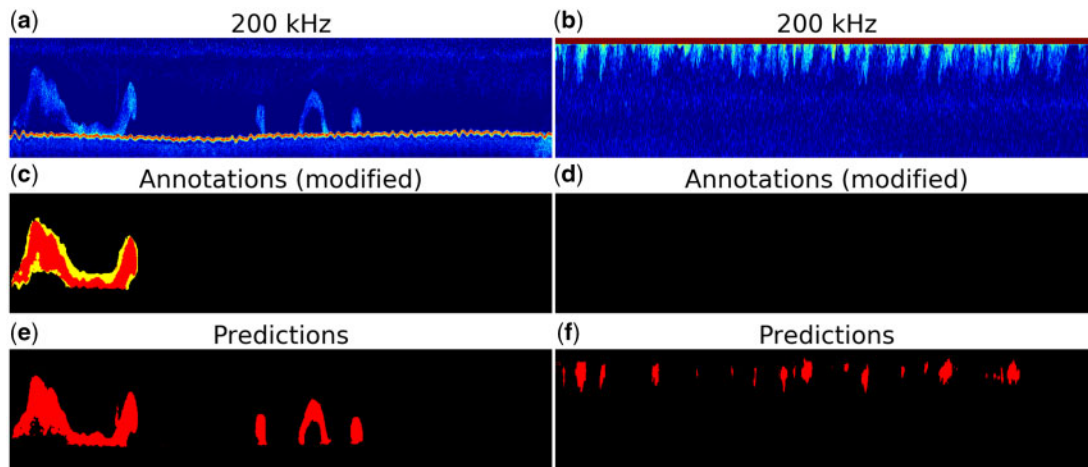


Figure 6. (a and b) The 200-kHz echograms. (c and d) Modified “sandeel” annotations in red (grey in print), the “ignore” pseudo class in yellow (light grey in print), and the “background” class in black. (e and f) Prediction of the “sandeel” class. (a, c, and e) Echogram with the absence of fish annotations in the right-hand side of the image. (b, d, and f) Echogram with false-positive predictions of sandeel close to the surface, possibly due to a zooplankton layer that the network is not trained to recognize. Axes are similar to Figure 1, where the vertical and horizontal axes represent depth and time, respectively.

positive predictions made by the model will be penalized when calculating the precision/recall curves. This illustrates a common problem encountered in the data, where image data are recorded with an abrupt absence of annotations (the remaining part of the echogram is annotated as “background”), c.f. the discussion for a possible explanation.

In some cases, the model makes false-positive predictions for “sandeel”. This is a common problem near the surface, where the model often classifies high s_v values, which could be caused by dense plankton layers, as “sandeel”. This class is annotated as “background” during training, but since we did not balance the training dataset for this case, as we did for the bottom, the model was not exposed to these “background” layers. The model has, consequently, not learned to annotate it as “background” and occasionally erroneously classifies them as “sandeel” instead.

Discussion

The objectives of this article were to define, train, and apply a deep CNN model that performs automatic classification of labelled multifrequency echosounder data and discuss how deep CNNs may be utilized for acoustic data. One of the main strengths of this model is that it does not require prior feature extraction steps, as it works directly on the output from the echosounder. These learned features may be both energetic and morphometric (Reid, 2000; Korneliussen, 2018), and there is no need to specify the features explicitly or to what degree one or the other should be used. The method also avoids any pixel averaging by school or region before applying the classifier, as the method works on high-resolution data. As with all neural networks, model interpretation is difficult. In its current design, the CNN does not provide information relating to feature importance, making it less transparent when compared to conventional methods (e.g. random forest) that work using hand-crafted features.

The manual annotations from survey data may be uncertain, and the uncertainty is not explicitly coded within the data. When using predefined features, the number of parameters in the classification model is typically lower than what is needed for CNNs. In those cases, it has been recommended to use a high-quality

training set where classifications are certain (Korneliussen *et al.*, 2016). Training a CNN requires a large amount of training data, and utilization of the full set of annotations from the survey may be needed. This has the drawback that low-quality annotation data may be used in training and validation but has the advantage that the data span the full variability across the survey. To some extent, we worked around this by adding the “possible sandeel” class to the “ignore” pseudo class. We recommend that future implementations use a combination of the above and assign a larger weight to annotations that have high certainty, e.g. those from a feature library (Korneliussen *et al.*, 2016), or allocate them to the test set only.

Using non-standard image data with annotations not made specifically for machine learning is a challenge. The annotations from the survey were designed for integrating sandeel backscattering intensity values, and assigning low s_v values to the sandeel class does not substantially contribute to the integrated sandeel backscatter. Consequently, using square bounding boxes that include background pixels does not substantially affect the integrated backscatter and is more efficient during manual annotation than drawing the school outlines. This represented a challenge in this study as the objective was to separate sandeel and background classes, and hence, refining the annotations was necessary. The modified annotations were important in making the method work. Modification of manually annotated acoustic observations may be a necessary step when using annotations to build automatic classification models such as CNNs.

Addressing the class imbalance by exposing the network to balanced mini batches of the data that contained all classes was necessary. The “other” and “sandeel” classes could be balanced, since they were annotated, but balancing the “background” class was more challenging. This class was a combination of seabed, plankton layers, empty water, and any other unknown scatterers. For the seabed, we solved this by balancing the training set with respect to crops close to seabed (since we had the bottom approximately detected), but we did not balance this for the unlabelled surface layer. This layer is most likely composed of near-surface phytoplankton blooms, specifically high densities of the gas

producing *Phaeocystis*, which produce high levels of acoustic backscattering intensity at 18 and 38 kHz. Since there is some overlap in the backscattering intensity of the surface plankton layer and of sandeel schools, the network would occasionally misclassify the “background” class as “sandeel”. A possible solution to this problem could have been to have implemented an unsupervised segmentation of the background class and then balance the training dataset based on the resulting classes. Consequently, addressing class imbalance is important for the *actual* classes in the data, not only for those that are annotated, and represents a general challenge when implementing supervised methods on acoustic data.

Processing the whole survey time series using the same automated algorithm is more efficient, consistent, and cost-effective than processing the data manually. We deliberately separated the training and test dataset by years to see if the network could generalize across years. The results showed that the performance changed by year, but this was not necessarily explained by the training and test datasets (Figure 5 and Supplementary Figure S2). The annotation issues noted above could account for parts of the discrepancies, but there were also other features that may have caused the network to perform differently across years. The annotation of sandeel schools is easier for large schools (due to more stable frequency responses and higher signal to noise), and school size tends to increase with high sandeel abundance (Johnsen *et al.*, 2017). In years with low sandeel abundance, a higher proportion of small schools cause a more uncertain categorization. Furthermore, weather condition may affect the schooling behaviour, which affects sandeel school detection.

When reviewing model performance by year, especially when including all background pixels (Supplementary Figure S2), some of the discrepancies may have arisen due to erroneous annotation. For survey years up to 2008, the labelling tool was under development and labelling was less efficient, and typically square annotation boxes were used. A bug in the annotation software was discovered in 2013 that led to incorrect storing of the annotation information (but not the exported backscattering intensity values). This may explain the improvement in performance in later years. For 2015, the weather conditions were rough, which led to underestimated biomass as stated in the 2015–2017 survey reports. For 2016, the “0-group sandeel” class was introduced due to large amounts of juveniles, which indicates a change in the system that may cause the model to perform differently, or alternatively, caused the labelling to be more challenging. From this perspective, reviewing the performance of the model across years is an efficient tool to identify any potential biases in the data series, but these considerations also apply to our benchmark.

There are several future directions in which we would like to take this. Further improvements of the model could be to include net sampling data and depth as separate inputs, where net samples could provide additional species information, and due to the conical shape of the echosounder beam, range could be used to compensate for range effects (e.g. that schools at short range look different at longer ranges). We would also want to include the uncertainty of the acoustic categorizing to the survey abundance estimates and, consequently, the fisheries advice when fully automating the annotation process. Another particularly interesting property of CNNs is transfer learning, i.e. that a network can be initialized from a previously trained network, and when presented with new data can update its weights. When a network is developed for sandeel classification, we can apply transfer learning and

adapt the network for different species, ideally across a wide range of surveys.

We have shown that a CNN can be reliably trained to categorize acoustic multifrequency observations. The main strength of this method is that the parameters can be learned directly from the echosounder output using manual labels as training data, i.e. there is no need to predefine features like frequency response, school morphology, etc., as the network learns the features directly from the training data. The method also allows us to code the tacit “knowledge” of a skilled operator, and it would be interesting to see if the method could be used to replicate different operators. In conjunction with more traditional, physics-based methods, this would enable us to study drift in expert judgements, explain annual differences, etc. When the network is trained on other surveys, we can transfer networks between surveys and look for differences in practices and test the implications. In our opinion, an end-to-end training approach opens possibilities not achievable when using conventional methods.

Supplementary data

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

Acknowledgements

This work is a part of the COGMAR project, funded by the Norwegian Research Council (grant 270966).

References

- Fallon, N. G., Fielding, S., and Fernandes, P. G. 2016. Classification of Southern Ocean krill and icefish echoes using random forests. *ICES Journal of Marine Science*, 73: 1998–2008.
- Foote, K. G. 1983. Linearity of fisheries acoustics, with addition theorems. *The Journal of the Acoustical Society of America*, 73: 1932–1940.
- Foote, K. G., Knudsen, H. P., Vestnes, G., MacLennan, D. N., and Simmonds, E. J. 1987. Calibration of acoustic instruments for fish density estimation: a practical guide. ICES Cooperative Research Report 144.
- Freeman, S., Mackinson, S., and Flatt, R. 2004. Diel patterns in the habitat utilisation of sandeels revealed using integrated acoustic surveys. *Journal of Experimental Marine Biology and Ecology*, 305: 141–154.
- Furness, R. W. 2002. Management implications of interactions between fisheries and sandeel-dependent seabirds and seals in the North Sea. *ICES Journal of Marine Science*, 59: 261–269.
- Gastauer, S., Fässler, S. M. M., O'Donnell, C., Høines, Å., Jakobsen, J. A., Krysov, A. I., Smith, L., *et al.* 2016. The distribution of blue whiting west of the British Isles and Ireland. *Fisheries Research*, 183: 32–43.
- Gjøsæter, H., Bogstad, B., Tjelmeland, S., and Subbey, S. 2015. A retrospective evaluation of the Barents Sea capelin management advice. *Marine Biology Research*, 11: 135–143.
- Haralabous, J., and Georgakarakos, S. 1996. Artificial neural networks as a tool for species identification of fish schools. *ICES Journal of Marine Science*, 53: 173–180.
- Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. 2015. Hypercolumns for object segmentation and fine-grained localization. *In* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 447–456.
- Horne, J. K. 2000. Acoustic approaches to remote species identification: a review. *Fisheries Oceanography*, 9: 356–371.
- ICES. 2016. Report of the Benchmark Workshop on Sandeel (WKSand 2016). ICES Document CM 2016/ACOM: 33.

- Ioffe, S., and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. <http://arxiv.org/abs/1502.03167> (last accessed 26 November 2019).
- Johnsen, E., Pedersen, R., and Ona, E. 2009. Size-dependent frequency response of sandeel schools. *ICES Journal of Marine Science*, 66: 1100–1105.
- Johnsen, E., Rieucan, G., Ona, E., and Skaret, G. 2017. Collective structures anchor massive schools of lesser sandeel to the seabed, increasing vulnerability to fishery. *Marine Ecology Progress Series*, 573: 229–236.
- Karp, W. A., and Walters, G. E. 1994. Survey assessment of semi-pelagic gadoids: the example of Walleye Pollock, *Theragra chalcogramma*, in the Eastern Bering Sea. *Marine Fisheries Review*, 56: 8–22.
- Kloser, R. J., Ryan, T., Sakov, P., Williams, A., and Koslow, J. A. 2002. Species identification in deep water using multiple acoustic frequencies. *Canadian Journal of Fisheries and Aquatic Sciences*, 59: 1065–1077.
- Korneliussen, R. J. (Ed). 2018. Acoustic target classification. ICES Cooperative Research Report 344. 104 pp.
- Korneliussen, R. J., Heggelund, Y., Macaulay, G. J., Patel, D., Johnsen, E., and Eliassen, I. K. 2016. Acoustic identification of marine species using a feature library. *Methods in Oceanography*, 17: 187–205.
- Korneliussen, R. J., and Ona, E. 2003. Synthetic echograms generated from the relative frequency response. *ICES Journal of Marine Science*, 60: 636–640.
- Long, J., Shelhamer, E., and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440. Boston, MA.
- Macer, C. T. 1966. Sand Eels (Ammodytidae) in the Southwestern North Sea: Their Biology and Fishery. *Fishery Investigations Series 2*, 24. H.M. Stationery Office, London.
- MacLennan, D. N., Fernandes, P. G., and Dalen, J. 2002. A consistent approach to definitions and symbols in fisheries acoustics. *ICES Journal of Marine Science*, 59: 365–369.
- 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.
- Mukai, T., and Amakasu, K. 2016. Comparison of the volume backscattering strength measured by EK60 and EK80. *The Journal of the Acoustical Society of America*, 140: 3242–3242.
- Ona, E. 2003. An expanded target-strength relationship for herring. *ICES Journal of Marine Science*, 60: 493–499.
- Proud, R., Mangeni-Sande, R., Kayanda, R. J., Nyamweya, C., Ongore, C., Everson, I., and Elison, I. 2020. Acoustic identification of schools of the Silver Cyprinid *Rastrineobola argentea* in Lake Victoria using Random Forests. *ICES Journal of Marine Science*, 77: 1379–1390.
- Reid, D. G. 2000. Report on echo trace classification. ICES Cooperative Research Report 238. International Council for the Exploration of the Sea, Copenhagen, Denmark.
- Ronneberger, O., Fischer, P., and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. <https://arxiv.org/abs/1505.04597> (last accessed 26 November 2019).
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning representations by back-propagating errors. *Nature*, 323: 533–536.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252.
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. 2013. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. <http://arxiv.org/abs/1312.6229> (last accessed 26 November 2019).
- Simmonds, J., and MacLennan, D. 2005. *Fisheries Acoustics: Theory and Practice*. Blackwell Science, Oxford. 437 pp.
- Trenkel, V. M., Mazauric, V., and Berger, L. 2008. The new fisheries multibeam echosounder ME70: description and expected contribution to fisheries research. *ICES Journal of Marine Science*, 65: 645–655.
- Winslade, P. 1974. Behavioural studies on the lesser sandeel *Ammodytes marinus* (Raitt) II. The effect of light intensity on activity. *Journal of Fish Biology*, 6: 577–586.
- Wuillez, M., Ressler, P. H., Wilson, C. D., and Horne, J. K. 2012. Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery. *The Journal of the Acoustical Society of America*, 131: EL184–EL190.
- Wright, P. J., Jensen, H., and Tuck, I. 2000. The influence of sediment type on the distribution of the lesser sandeel, *Ammodytes marinus*. *Journal of Sea Research*, 44: 243–256.

Handling editor: David Demer