ICES Journal of Marine Science, 2025, Vol. 82, Issue 5, fsaf069
https://doi.org/10.1093/icesjms/fsaf069
Received: 3 February 2025; **revised:** 28 April 2025; **accepted:** 2 May 2025
Original Article

Integrating and assessing machine learning acoustic target classification models for fish survey estimations

Nils Olav Handegard [1,*], Arne Johannes Holmin [2], Ahmet Pala [3], Ingrid Utseth [4], Espen Johnsen [2]

[1]Department of Acoustics and Observation Methodologies, Institute of Marine Research, Nordnesgaten 50, Bergen 5005, Norway
[2]Department of Pelagic Fish, Institute of Marine Research, Nordnesgaten 50, Bergen 5005, Norway
[3]Department of Mathematics, University of Bergen, Allegaten 41, 5007 Bergen, Norway
[4]Department of Image Analysis, Machine Learning and Earth Observation, Norwegian Computing Center, Gaustadalleen 23A, 0373 Oslo, Norway

*Corresponding author. Acoustics and Observation Methodologies, Institute of Marine Research, P. O . Box 1870 Nordnes, 5817 Bergen, Norway. E-mail: nilsolav@hi.no

Abstract

Scientific acoustic-trawl surveys collect data that are used to track fish and zooplankton populations over time. Most rely on manual annotation during acoustic target classification, but automated methods have been proposed. Here, we report on a framework for testing deep learning-based acoustic classification models and integrating them into the survey estimation process. The approach was applied to North Sea lesser sandeel (*Ammodytes marinus*) surveys from 2009 to 2024. Three U-Net-based models were tested: a baseline model, a depth-aware model, and a model trained with similarity-based sampling for the foreground class. A threshold based on the training years was applied to the models' SoftMax outputs. The official sandeel estimation process was used as a starting point, replacing input data with model predictions. The biomass estimates were generally similar between manual annotations and model-based estimates, but variation existed across years. The baseline model misclassified a surface layer as sandeel and was prone to bottom contamination, causing larger deviations from official estimates. Discrepancies between the similarity-based model and the official estimates resulted from an incorrectly applied SoftMax threshold, leading to missing school interiors and indicating threshold sensitivity. Unlike traditional F1 score evaluations commonly used in image-based classification, our comparison assessed predictions in a survey-relevant context. The evaluation indicated that full automation was not yet feasible, but the predictions could be used as starting points for manual scrutiny. Annotating a subset of the data to refine thresholds or employing more advanced active learning approaches could enhance efficiency. These methods could enable faster, more consistent survey annotation.

Keywords: acoustic-trawl survey; acoustic target classification; multi-frequency echograms; semantic segmentation; big data; deep learning; fisheries management; StoX

Introduction

Scientific acoustic-trawl surveys (MacLennan and Simmonds 2005) are used to monitor changes in abundance, distribution, and population structure of fish and zooplankton (Gunderson 1993). Annual survey estimates are used as input to fisheries assessment models and provide time series data on changes in fish and zooplankton populations. They are an important complement to fisheries data in fish stock assessment models. An acoustic-trawl survey consists of collecting acoustic data, usually along transects, biological sampling, annotating the acoustic data with an acoustic category, and integrating the backscatter by category over a transect. Nets or trawls are typically used for biological sampling, where information on species, length groups, and other biological parameters such as age are recorded and combined with the acoustic data to estimate abundance or biomass.

Most acoustic trawl-surveys rely on manual annotation during the acoustic target classification (ATC) step. This is usually achieved through manually scrutinizing the data using desktop applications like Echoview (Myriax, Australia), LSSS (Korneliussen et al. 2016), or similar. During the process, the acoustic backscatter is assigned to a category representing a species or a group of species. The approach attributes all or a proportion of the total backscatter over a region to one or several acoustic categories. The process involves defining the seabed, removing noise, adjusting the acoustic density thresholds, drawing the region outlines, and attributing the acoustic backscatter to an acoustic category. In some surveys, the species and size distribution sampled from the biological samples is also used when attributing the acoustic backscatter. When multiple frequency channels are present, the frequency response may be used to aid the process.

ATC has long been considered a central challenge in fisheries acoustics (MacLennan and Holliday 1996), and several approaches have been taken to automate the process (Korneliussen 2018). Features from schools, elementary distance sampling unit features, and features across strata or regions can be used for classification (Reid 2000, Reid et al. 2000), together with machine learning (ML) classification algorithms (Haralabous and Georgakarakos 1996). The use of the relative frequency response between different echosounder channels is commonly used (Kloser et al. 2002, Korneliussen and Ona 2003), or by combining the relative frequency re-

sponse and morphological parameters (Korneliussen et al. 2009, Komiyama et al. 2024).

Numerous ML approaches have been proposed to automate the ATC. Rezvanifar et al. (2019) introduced a framework combining a region of interest extractor with a deep learning-based image classifier. Similar methods detect regions of interest before applying classification methods, while Marques et al. (2021b) proposed integrating these steps using end-to-end deep learning frameworks such as Faster R-CNN (Ren et al. 2015) and YOLOv2 (Redmon and Farhadi 2017). End-to-end learning refers to training a model directly from input data to output predictions, bypassing traditional feature extraction and stepwise data processing. Notable examples include the introduction of deep convolutional neural networks (CNNs) for digit recognition (Lecun et al. 1998) and the breakthrough in image classification on the ImageNet dataset (Krizhevsky et al. 2012). Training these models typically requires extensive labelled data, though Choi et al. (2021) introduced a semi-supervised approach and Pala et al. (2024) used a self supervised approach to reduce this dependency. Marques et al. (2021a) presented an instance segmentation framework for precise bounding box detection in herring schools.

Another important category of models is semantic segmentation, which predicts acoustic categories for each sample in an echogram. Brautaset et al. (2020) utilized a U-Net-based approach (Ronneberger et al. 2015), further analysed by Ordoñez et al. (2022) with different resolutions and depth information. A further improvement was achieved by Pala et al. (2023) by introducing a similarity-based sampling method. Vohra et al. (2023) compared Attention U-Net, U-Net, and DeepLabV3, while Choi et al. (2023) combined semantic segmentation with unsupervised learning.

The ML models by Brautaset et al. (2020), Ordoñez et al. (2022), Choi et al. (2021, 2023), and Pala et al. (2023) were trained on labelled survey data on Norwegian lesser sandeel (*Ammodytes marinus*) (hereafter sandeel). The sandeel is a small, swim bladder-less fish that spends much of its life burrowed in sandy seabeds with low silt and clay content (Macer 1966, Wright et al. 2000). In spring, adults emerge from the sand at dawn to form pelagic schools and feed on zooplankton (Winslade 1974, Freeman et al. 2004, Johnsen et al. 2017). The sandeel is vital prey for seabirds, seals, larger fish (Furness 2002), and supports commercial fisheries.

Most of the ML models applied to ATC are taken from the image classification domain and are trained end-to-end to predict the acoustic categories. However, these models are usually evaluated using image classification metrics, such as F1 scores, without considering the underlying backscatter distribution. Although an F1 score offers a robust test for image-based classification capabilities, it lacks the link to the use of the data and its usefulness to provide robust estimates of fish abundance, which requires integrating and aggregating the acoustic backscatter intensities associated with the foreground class. As an example, if an echogram 'pixel' with a low backscattering intensity is assigned to the wrong class, the impact on the result is much less than if the associated backscattering intensity is high. So far, most models have used the F1 score as the test criteria, and it can be argued that they are not truly end-to-end since they are tested on an intermediate step.

The objective of this paper is to establish a framework for testing predictions for automated methods for ATC, where the prediction results are integrated into the entire survey estimation process used for acoustic trawl surveys. To evaluate their performance, the survey estimate and integrated acoustic backscatter transect values are compared with those obtained using traditional standard manual ATC methods. The approach will be demonstrated on the North Sea lesser sandeel (*A. marinus*) acoustic-trawl survey time series.

Materials and methods

The sandeel survey

Since 2005, the Institute of Marine Research in Norway has conducted acoustic trawl surveys in northeastern North Sea sandeel areas (Johnsen et al. 2017). The 2007–2024 surveys used various vessels, including RV Johan Hjorth, RV GO Sars, FV Brennholm, FV Eros, FV Kings Bay, and RV Kristine Bonnevie (Table 1).

The main objectives of the 2007 and 2008 surveys were methodology development and spatial density mapping of sandeel. From 2009 onward, the survey was used to produce a time series for assessing sandeel stock status. The sandeel grounds were divided into several strata, which were covered using zigzag (Strindberg and Buckland 2004) or parallel transects with a random starting point in each stratum (Fig. 1). As sandeel burrow into the sand at night, the transects were only covered during daylight hours. A Campelen 1800 demersal trawl (Nguyen et al. 2015) was used for trawl sampling. The trawl opening height was ∼4 m, the door distance was ∼50 m, and the trawl was equipped with a 10 mm-meshed cod-end. The trawl targeted sandeel schools during the daytime at a trawling speed of 3 knots. A sandeel dredge with a width of 1 m and net mesh size of 5 mm (Johnsen and Harbitz 2013) at a towing speed of 2 knots was used to catch sandeel burrowed in the seabed. The vessels were equipped with Simrad EK60/80 echosounders (Kongbserg Discovery, Norway) operating at 18, 38, 120, and 200 kHz frequency channels, except FV Brennholm, where the 120 kHz transducer was not available (Table 1). Standard calibration procedures (Demer et al. 2015) ensured accurate data, with echosounders set to a 1.024 ms pulse duration, 3–4 Hz ping repetition frequencies, and a vessel speed of ∼10 knots when covering the transects. Further details can be found in Johnsen et al. (2009) and Komiyama et al. (2024).

During the surveys, acoustic data were categorized as 'sandeel', 'other', '0-group sandeel', or 'possible sandeel'. The classes 'sandeel' and 'other' have been used annually, with 'sandeel' being the only one included in estimates used for the advisory process. The 'possible sandeel' category was introduced for schools with uncertain frequency responses, and '0-group sandeel' category was added in 2016 due to an unusually high juvenile density. School sizes ranged from a few meters to over 1 km, spanning much of the water column (Johnsen et al. 2017). During manual scrutiny using LSSS, the contouring of sandeel schools was based on the 200-kHz data due to its strong sandeel signal. Acoustic categorization, which underpinned the classification process, was based on the frequency response of the volume backscattering coefficient (s_v), defined as the mean backscattering intensity per cubic meter (m^3) (c.f., Johnsen et al. 2009). For echo integration, the 38 kHz frequency channel was used. After annotation, the corresponding acoustic data were stored in an internal database in the LSSS system and exported

Table 1. Overview of the sandeel survey series.

Year	Vessel	Operating frequencies (kHz)	Echosounder
2007	RV Johan Hjort	38, 18, 120, 200	EK60
2008	RV Johan Hjort	38, 18, 120, 200	EK60
2009	RV G.O. Sars	38, 18, 70, 120, 200, 333	EK60
2010	RV Johan Hjort	38, 18, 120, 200	EK60
2011	RV Johan Hjort	38, 18, 120, 200	EK60
2012	FV Brennholm	38, 18, 200, 333	EK60
2013	FV Eros	38, 18, 70, 120, 200	EK60
2014	FV Eros	38, 18, 70, 120, 200, 333	EK60
2015	FV Eros	38, 18, 70, 120, 200, 333	EK60
2016	FV Eros	38, 18, 70, 120, 200, 333	EK60
2017	FV Eros	38, 18, 70, 120, 200, 333	EK60
2018	FV Eros	38, 18, 70, 120, 200, 333	EK60
2019	FV Eros	38, 18, 70, 120, 200, 333	EK60
2020	FV Eros	38, 18, 70, 120, 200, 333	EK80
2021	FV Kings Bay	38, 18, 70, 120, 200,	EK80
2022	RV Johan Hjort	38, 18, 120, 200, 333	EK80
2022	RV Kristine Bonnevie	38, 18, 120, 200	EK80
2023	RV Kristine Bonnevie	38, 18, 120, 200	EK80
2024	RV Johan Hjort	38, 18, 120, 200, 333	EK80

The survey has been carried out using a combination of chartered fishing vessels (FV) and IMR research vessels (RV).

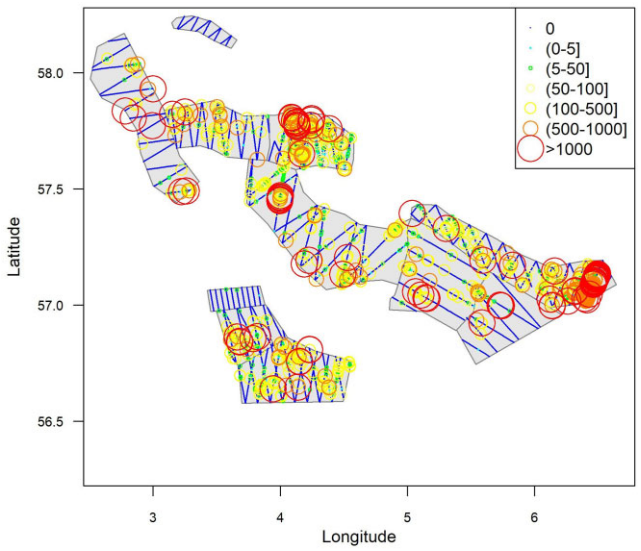

Figure 1. Survey design for the 2020 survey with strata (grey polygons) and transects, where the geographical distribution of the sandeel nautical area scattering coefficient (NASC) values are presented per 0.1 nautical mile along-track segments. The size of the circles indicate NASC, while zero values are shown as dots (from Johnsen and Kvamme 2024).

in a depth and log-distance grid of 10 m and 0.1 nautical miles.

For the EK60 data, the range resolution was determined by the pulse length. This allowed all the data to be stacked in a three-dimensional grid in range, time, and frequency channels. If, for some reason, another pulse length was used, the data were interpolated onto the same range resolution as the standard 1.024 ms pulse duration data. For the EK80 implementation, range resolution depended on the operating frequency. To align the data with the EK60 grid, it was resampled to the same dimensions using a weighted mean, where the weights reflected the degree of overlap with the original grid.

Table 2. The different models used in this study for classification.

Model	Description	Citation
1	Original labels	NA
2	Baseline U-Net implementation	Brautaset et al. (2020)
3	The U-Net including auxiliary data	Ordoñez et al. (2022)
4	Balanced training method	Pala et al. (2023)

Machine learning models for ATC

The U-Net based algorithms available for predicting pixel-level annotations were used in this study (Table 2). While the same model architectures and acoustic data were employed as in the cited studies, the original models had been trained on datasets pre-processed using a different pipeline than the one applied here. However, the underlying acoustic data and model algorithms remained the same. The prediction step used the same input data across all models.

Raw acoustic data (.raw files) and annotation data from the LSSS system (.work files) were converted to the Zarr format, a cloud-friendly format that supports random access and seamless integration with standard software packages. Using Python, this format can be directly accessed through Xarray objects, enabling easy interaction with both raw data and annotations. The s_v data preprocessor gridded the raw data into an N-dimensional array, which was stored as a Zarr dataset. Annotations from LSSS work files were reformatted to align with the same grid as the s_v files, while grid-independent annotations were stored in a Parquet file (Fig. 2a). See Supplementary Material for details.

The baseline model was initially designed as a CNN based U-Net model (Ronneberger et al. 2015) to perform pixel-wise segmentation on echogram data for sandeel detection (Brautaset et al. 2020). The baseline model used a weighted loss function and a class-based sampling strategy to address the heavy class imbalance in the dataset. The output was the pixelwise SoftMax maps for the sandeel class for each pixel (Fig. 2b), c.f. Brautaset et al. (2020) for details. The model was trained on data from the 2011, 2013–2016 surveys, and the model was tested on the 2007–2010, 2017, and 2018 surveys.

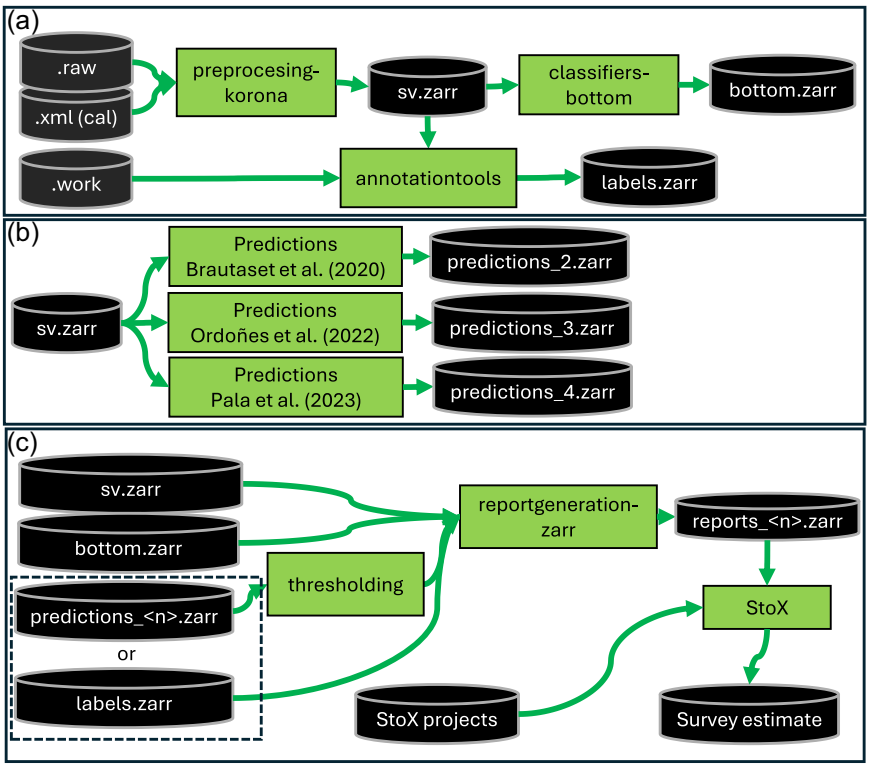

Figure 2. The processing steps from raw data to survey estimates. (a) The preprocessing and regridding of the acoustic data (.raw files) and calibration data (.xml) through the 'preprocessing-korona' module, the conversion of annotation data from the LSSS system (.work files) through the 'annotationtools' module, and the simple bottom detection algorithm through the 'classifiers-bottom' module. (b) The volume backscattering coefficient (s_v) data are used as an input to the machine learning (ML) models, and the predictions contain the softmax output from the models. (c) The s_v data and the bottom are combined with the predictions from the ML models. The s_v data are masked by the above-bottom samples and the sandeel predictions, regridded on a 10 m by 0.1 nmi grid, imported to the StoX estimation program where the original input data is replaced, and the survey estimate is exported. See Supplementary Material for details.

The second model was built on the baseline model by adding auxiliary information to the network (Ordoñez et al. 2022). This model employed a broader data augmentation pipeline during training to increase robustness under diverse survey conditions. It also aimed to evaluate the effectiveness of different preprocessing strategies. The model architecture introduced additional layers that improved accuracy and resulted in higher F1 scores than the baseline model. The output was the pixelwise SoftMax maps for the sandeel class for each pixel (Fig. 2b), c.f. Ordoñez et al. (2022) for details. The training and testing data were the same as for the baseline model.

The third model (Pala et al. 2023) focused on the class imbalance when training models on acoustic data, where background pixels far outnumbered sandeel pixels. This model was built upon the baseline model by employing a similarity-based sampling strategy. This technique selectively increased the representation of sandeel-like pixels during training, allowing the model to better learn sandeel-specific characteristics without being overwhelmed by the background class. Training on a more balanced representation of sandeel and background pixels, the model achieved improvements in segmentation precision and was more effective in difficult detection regions. As with previous models, the output is a pixelwise SoftMax map (Fig. 2b), c.f. Pala et al. (2023) for details. The training and testing data were the same as for the baseline model.

The models predicted the SoftMax output for each sample in the echogram for the sandeel class. To convert the SoftMax labels to binary labels similar to the data from the manual ATC, a threshold was applied to the SoftMax predictions (Fig. 2c). However, the same threshold cannot be applied across all models, and we used the following strategy for setting the thresholds: For each model, we calculated the threshold value that maximized the F1 score for each *training* survey year (2011, 2013–2016). While using the threshold that maximizes the F1 score for each individual year would improve performance, this approach is not a fair evaluation metric because calculating the F1 score requires access to ground-truth labels. We used the median threshold for each model for the training years when thresholding the data for all years during prediction (Fig. 3). The thresholds were applied to the SoftMax outputs to classify pixels as sandeel or non-sandeel.

The CNN models are not aware of the seabed. During the training, the seabed is treated as the background class, and the models can discriminate between the seabed and the foreground class. By balancing the bottom samples of the background class during the training, the model's performance can be further improved to avoid predicting seabed as the foreground class. However, samples below the seabed, e.g. the second echoes from sandeel schools, are occasionally predicted as sandeel. To avoid these being included in the estimates, we ran a simple bottom detection algorithm and removed

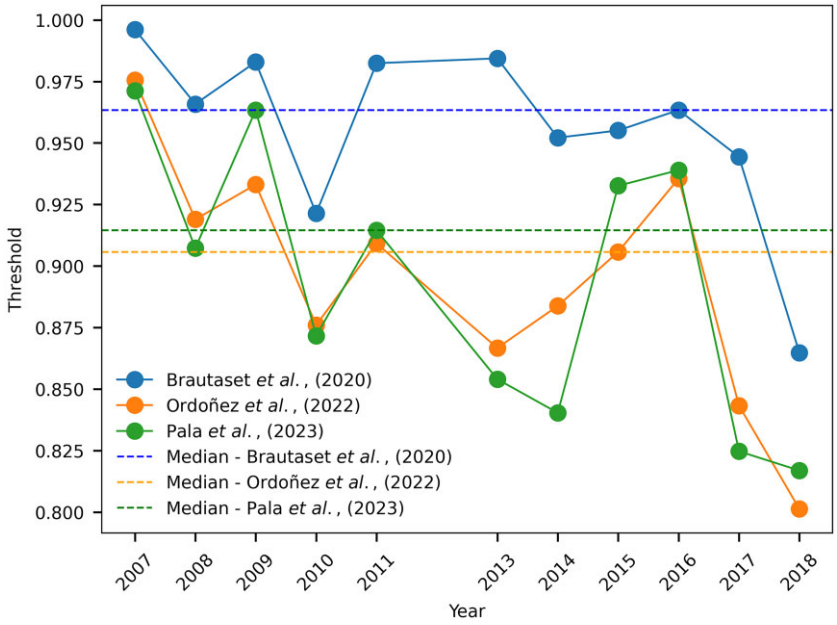

Figure 3. The threshold values that maximized the F1 scores for each machine learning model and each survey year used for training and testing. The median threshold values for the training years that maximized the F1 score for each model are shown as horizontal dashed lines. We used these values for all testing years, including the years after 2018.

the predictions in a buffer 10 pixels *below* the seabed. This step was not intended to remove the seabed itself, but rather to eliminate second echoes from sandeel detected below the seabed.

Integration and survey estimation

For each ping, the s_v data, the (binary) bottom predictions b and the (binary) ML based sandeel predictions m were combined to provide the estimates of the nautical area scattering coefficient (NASC, mean backscattering intensity over a given depth range per nautical mile squared),

$$s_{A,i} = 4\pi \left(1852^2\right) \int_{z_i}^{z_i+10} s_v(z) \cdot m(z) \cdot b(z)\, dz,$$

where z is the range from the transducer, z_i is the range in 10 m layers, and i is the layer number. Note that we used range instead of depth which is the standard procedure. The depth conversion requires regridding, and we chose to stay in range domain. The s_A values were then averaged over 0.1 nmi distance bins and only data recorded along the survey transects were included in the analyses. The time and geographical position for the first and last ping was calculated for each bin, which provided a structured dataset that enabled integrating the machine learning-based classification models with the established fisheries abundance estimation workflows. See Supplementary Material for details.

The standard survey estimates of sandeel are considered to reflect the absolute abundance and biomass of sandeel in the survey areas (Johnsen and Kvamme 2024). Survey estimation follows standard procedures using the StoX software (Johnsen et al. 2019), in which each transect is defined and manually assigned to a stratum. Acoustic data from transits between transects and during trawling operations are excluded. For each transect, a set of biological stations is assigned. The mean vertically integrated s_A values are calculated per transect, and the acoustic energy is converted to area number density using a target strength relationship of $\mathrm{TS}_{38\,\mathrm{kHz}} = 20\log_{10}(L) - 93$ ($dB\ re\ 1\,\mathrm{m}^2$). Length distributions are derived from the assigned trawl stations. The resulting densities are averaged over geographical strata and multiplied by stratum area to estimate abundance by length group. Biomass is calculated as the product of abundance and individual weights. Precision is estimated by bootstrapping with 1000 iterations, resampling transects and trawl hauls with replacement within each stratum. The bootstrap summary provides estimates of the mean, confidence intervals and coefficient of variation.

In this study, we used the survey estimation procedure described above (Johnsen and Kvamme 2024) but replaced the NASC values with those derived from the ML-based ATC reports (c.f. Fig. 2c), while keeping the assigned strata and biological samples unchanged. The primary sampling units (PSUs) from the ML-based acoustic data were adjusted to match the existing PSUs, correcting for any discrepancies in the start and end times of the 0.1 nmi distance bins. For each case and year, the distance-weighted average NASC value per PSU and the corresponding biomass with a 90% confidence interval were calculated.

Results

Overall, the biomass estimates derived from manual label-based annotations and ML predictions were similar, but results varied between years (Fig. 4). When the models were developed, the data from 2019 and onward were not available and were not used for training or testing any of the models. The baseline ML model (Brautaset et al. 2020) exhibited the largest deviations from the official estimates, especially for 2009 and 2024. The baseline model is vulnerable to erroneously allocating surface layers of zooplankton

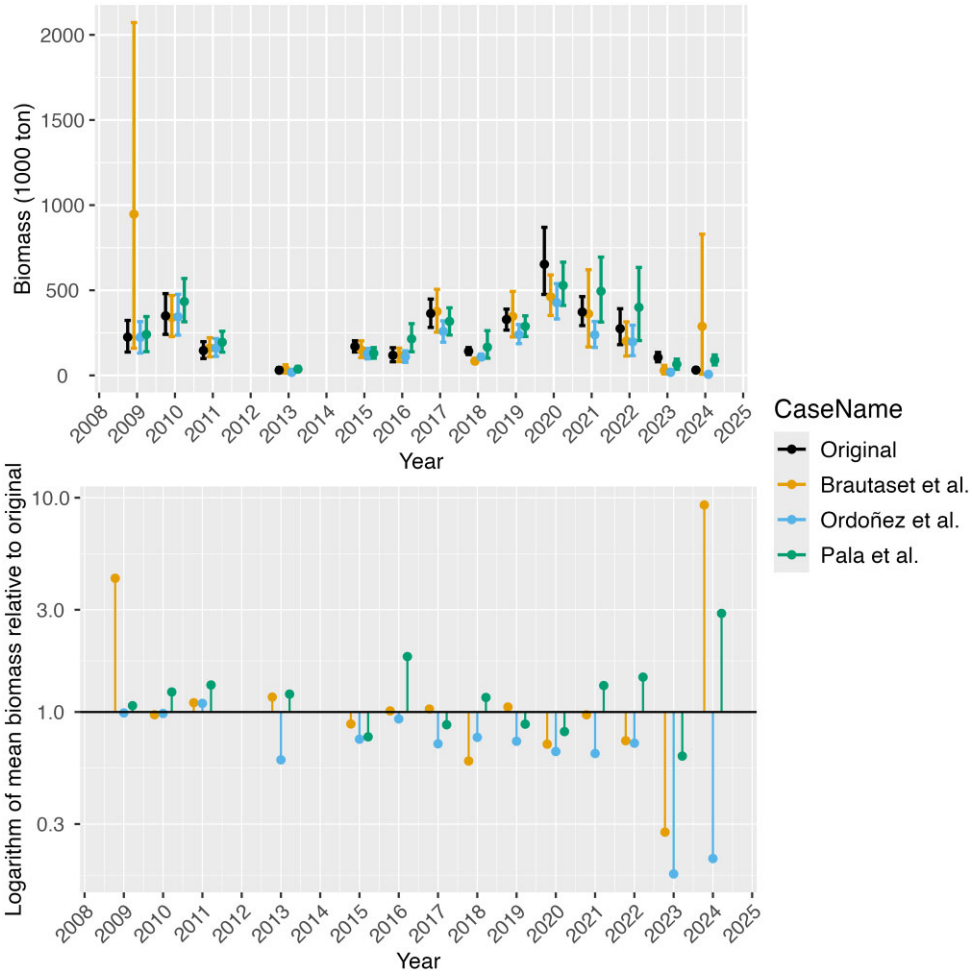

Figure 4. Upper panel: Total biomass of sandeel (age 1+) with a 90% confidence interval (5%–95%) for all survey areas combined per year, estimated from the acoustic sandeel surveys (Johnsen and Kvamme 2024). The original is the official estimate, whereas the three others are the estimates from Brautaset et al. (2020), Ordoñez et al. (2022), and Pala et al. (2023), respectively. Note that 2012 is missing since the data set lacked the 120 kHz echosounder data. The 2014 data failed due to missing data. Lower panel: the absolute difference between the prediction point estimate and the original estimate.

to sandeel as well as bottom contamination. The model that is depth aware (Ordoñez et al. 2022) performed better, and there were no major deviances between the original manual label-based estimate and the biomass estimates based on this model. The model based on the similarity-based training (Pala et al. 2023) was similar but was less robust for the later years, e.g. 2021 and 2022. Integration over range instead of depth did not substantially affect the estimates, as values for the training years remained consistent with the original results (Fig. 4).

To examine the performance of the three models in more detail, we analysed the 2019 survey estimates. All models performed reasonably well that year, and it was the first year that was not seen by the model (or modellers) at all. The PSU used in the estimation was the average across a transect, and there was reasonable agreement between the model predictions and labels (Fig. 5). One exception was with the predictions from Brautaset et al. (2020), which produced substantially higher NASC values than those derived from the manual labelling. We also compared the values on the finer 0.1 nmi resolution (Fig. 6). This discrepancy between the Brautaset et al. (2020) model and the labels appeared to originate from multiple data points, rather than a single location, as would be expected if only a few bottom pixels had been mistakenly classified as sandeel.

To understand what caused the discrepancies, we listed the 0.1 nmi PSUs that had the largest discrepancy between the integrated backscatter for sandeel over 0.1 nmi, masked by the original annotations and model predictions, respectively. In 2019, the PSU with the largest discrepancy between predictions and labels originated from the baseline model, due to misclassification of a surface plankton layer as sandeel (Fig. 7c). The case with the largest discrepancy for the Pala et al. model was the erroneous application of a low SoftMax threshold, where larger parts of the interior school were missing from the predictions (Fig. 7d).

Discussion

We developed a framework to evaluate ML model predictions against survey estimates, exemplified by the sandeel survey in the North Sea. We have used the predictions together with the standard method and software used for the survey (Johnsen et al. 2019), and thus moved beyond the common F1 score

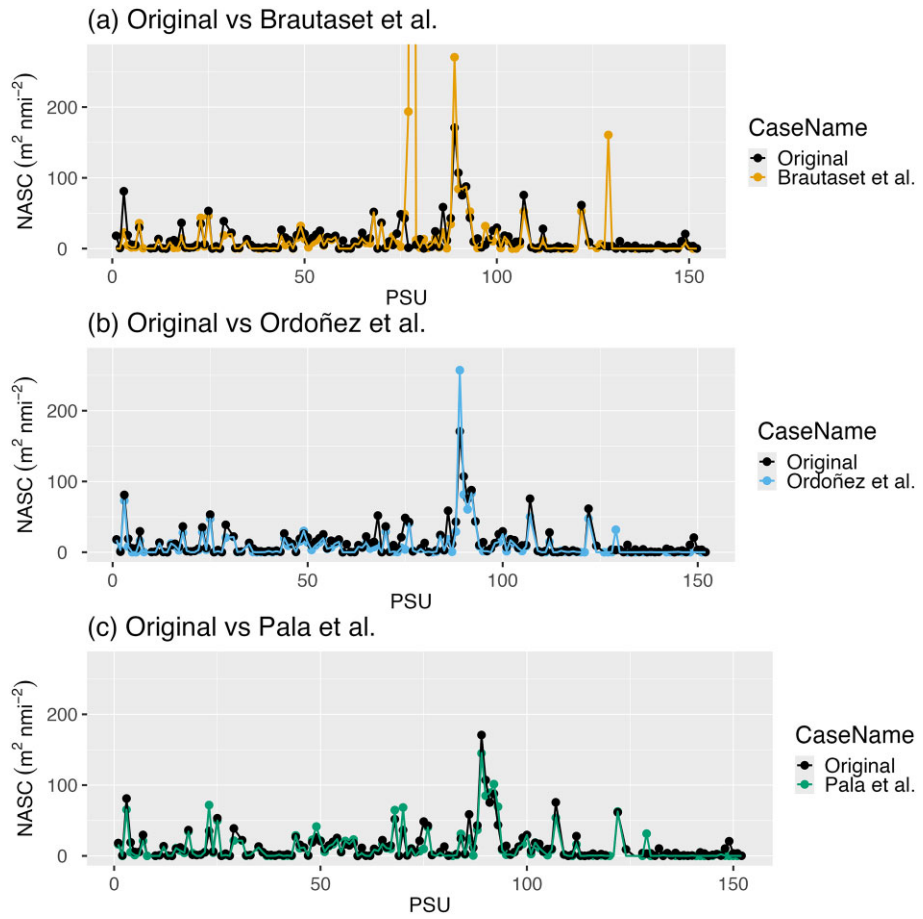

Figure 5. (a–c) Comparisons per PSU for the 2019 survey between the labels and the model predictions from Brautaset et al. (2020), Ordoñes et al. (2022), and Pala et al. (2023), respectively. Note that there is one outlier in the NASC values in the upper panel with a value of ∼2700 m^2 nmi^{-2}.

that is commonly used for evaluating the performance of image based methods applied to ATC. This allowed us to evaluate predictions using a metric that reflects the data's intended application.

Survey estimates based on the ML models showed lower variability and greater similarity to the original labelled data during the training years, while performance declined in more recent, unseen years. It is not necessarily surprising that we experienced a decline in performance for recent surveys. CNNs are sensitive to data shifts (model drift), which may result from factors such as weather conditions, changes in population structure (e.g. smaller schools, juvenile prevalence), species composition, or survey hardware. This emphasizes the importance of continuously monitoring model performance and to retrain networks as new data becomes available.

Cases with large discrepancies between predictions and labels in terms of acoustic backscatter were visualized and assessed. In general, the examples where the discrepancies were large are similar to the discrepancies reported earlier when inspecting where the F1 scores showed inferior performance (Brautaset et al. 2020). The model by Brautaset et al. (2020) tended to erroneously predict surface layers as sandeel. This caused large overestimations in terms of biomass for some years and is the cause of the large discrepancy in 2019 (Fig. 5). The model by Ordoñez et al. (2022) used the depth as input and performed better, but occasionally underpredicted the sandeel schools. The model by Pala et al. (2023) also partially misclassified surface layers as sandeel, but the largest discrepancy was caused by using a non-optimal threshold, emphasizing the need to reconsider the thresholding process. In a few cases, bottom signals were misclassified as sandeel. Such errors are difficult to detect using F1 scores alone, since only a few mislabelled samples will cause a large discrepancy in the integrated backscatter. However, these are relatively easy to identify when plotting the integrated values at a ping-to-ping resolution.

We used a hard threshold for translating the SoftMax output from the U-Net models to ML predictions, and the thresholds were set by the median of the thresholds that optimized the F1 score across the years up until and including 2018, for all three models, respectively. The optimal threshold varied between years, and a fixed threshold will cause performance to vary across years. This can be caused by changes in the fish abundance and distribution, which could affect the data distribution and, in turn, model performance and the optimal threshold. However, since labels are needed for setting a threshold for optimizing the F1 score, we cannot use this approach when predicting on survey data. One approach is to apply weights to the s_v data according to prediction confidence or to use a soft thresholding strategy. Another solution would be to set the threshold based on a human in the loop during the survey. After manually adjust-

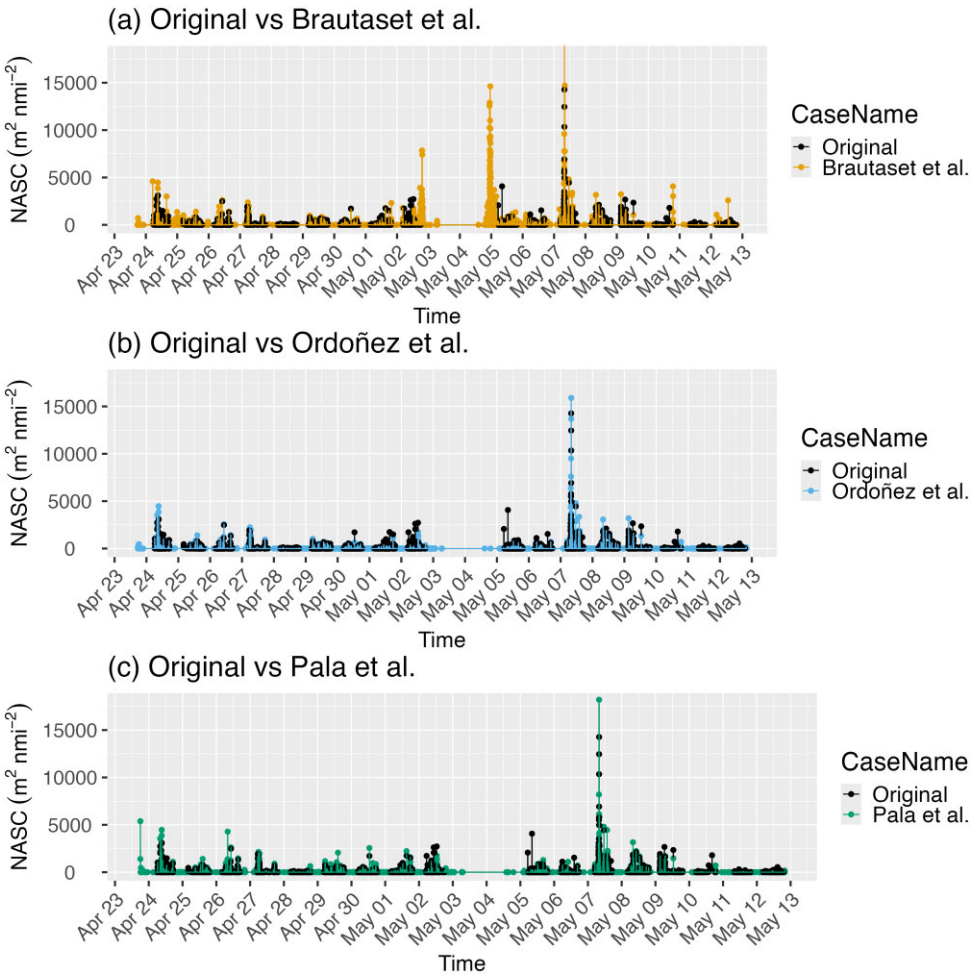

Figure 6. (a–c) Comparisons per 0.1 nmi for the 2019 survey between the labels and the model predictions from Brautaset et al. (2020), Ordoñes et al. (2022), and Pala et al. (2023), respectively.

ing the predictions from parts of the survey, the threshold could be updated and employed across all the PSUs. Instead of setting the threshold to minimize the F1 score, we could also set the threshold to minimize the error in the abundance estimate.

While adjusting the thresholds and retraining the model on recent data could help mitigate some performance loss, there is still a need to check and improve the performance during a survey. This can be achieved by using the predictions as starting points for a manual scrutiny process or annotating a subset of the data for adjusting the threshold. More sophisticated active learning approaches should also be considered (e.g. Budd et al. 2021) to select samples in the data to be annotated by a human and included in a training set. One such approach is to annotate the samples where the model's predictions are the least certain, based on a measure of the model's confidence. All these approaches could enable faster and more consistent processing of entire surveys, substantially reducing the time required for manual annotation.

Our approach linked all processing steps from raw data to biomass estimates (as illustrated in Fig. 2). In addition to testing the effect of ATC algorithms, the approach can also be used to test the effect of changes in other processing steps, if applicable. For instance, if a new bottom detection algorithm or noise detection algorithm is developed, the predictions can be used to create Boolean masks, which can then be combined with the classification predictions. If the processing step alters the backscatter values instead of the masks, such as correcting the backscatter data for noise (De Robertis and Higginbottom 2007, their Eq. 8) or transducer motion (Dunford 2005), the filtered backscatter data can be substituted while keeping the other parts unchanged. After replacing the combined mask or the s_v values, the effect on the survey estimate can be evaluated throughout the time series, rather than relying on a few test data sets, which is typically common practice when developing and testing algorithms.

This study presented a framework for evaluating ML model predictions in the context of survey-based abundance estimates. Future research could investigate dynamic or adaptive thresholding methods and incorporate additional data into the models, such as trawl samples, environmental information, and location data. Although this study concentrated on the lesser sandeel, the proposed framework can be applied to other species and types of surveys as long as the survey estimation step can be scripted. By modifying the input data and classification targets, similar workflows could facilitate automated analysis across various fisheries acoustics scenarios, including multispecies and mesopelagic surveys.

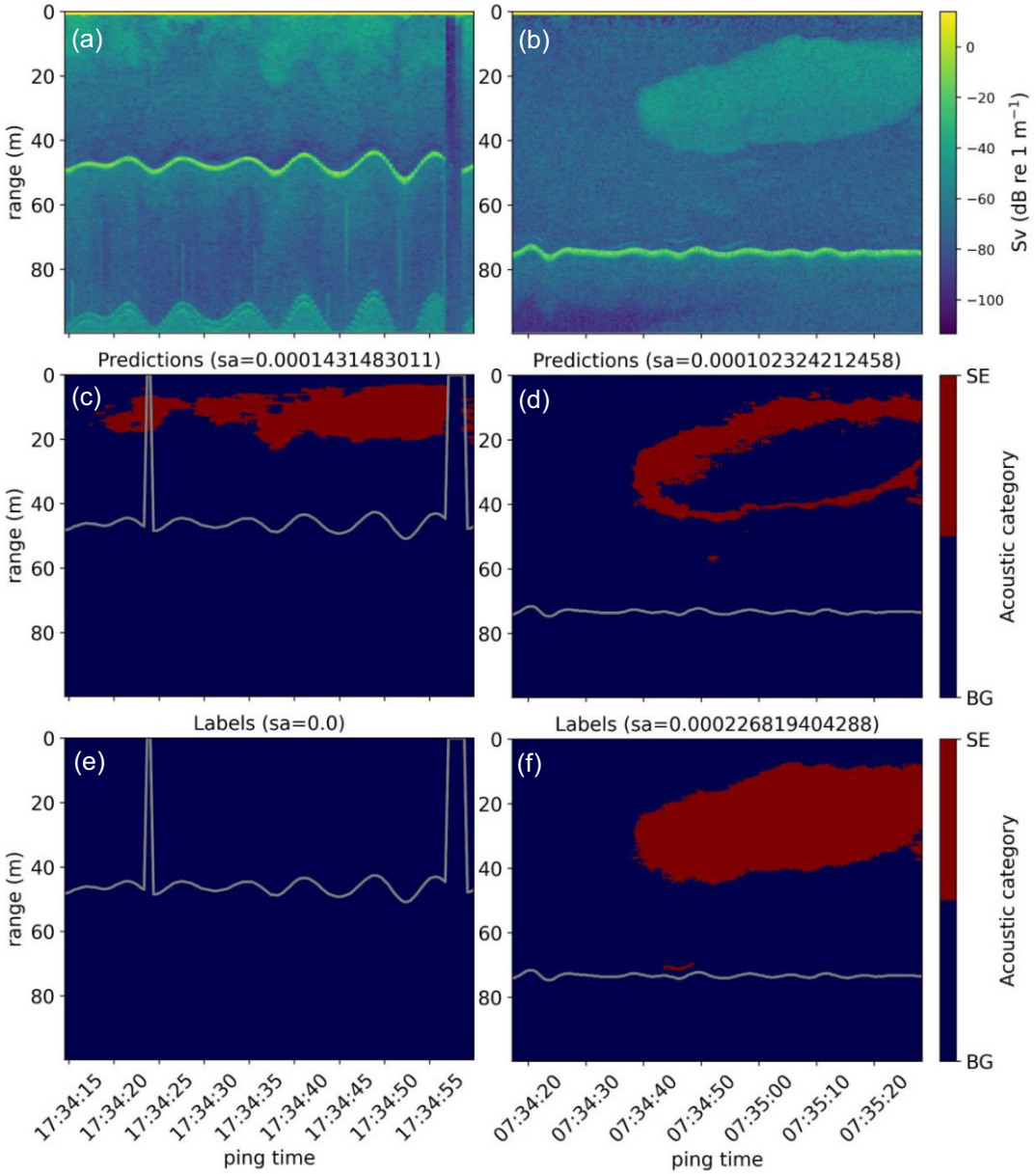

Figure 7. The (a and b) backscatter at 200 kHz for the 2019 survey for the 0.1 nmi with the largest discrepancy between the model predictions (c and d) and labels (e and f), for Brautaset et al. (2020) and Pala et al. (2023), respectively. The first example demonstrates the failure of the model to correctly assign the surface layer to the background (BG) category (c), whereas the second example shows the effect of choosing a threshold that is not tuned for the case resulting in the sandeel (SE) acoustic category to be erroneously assigned to the BG acoustic category (d). The white lines are the prediction from the bottom detection algorithm, and there are cases where the bottom detection fails.

Author contributions

N.O.H.: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. A.J.H.: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – review & editing. A.P: Formal analysis, Methodology, Software, Validation, Visualization, Writing – review & editing. I.U.: Formal analysis, Methodology, Software, Writing – review & editing. E.J.: Conceptualization, Investigation, Methodology, Resources, Validation, Writing – review & editing.

Supplementary material

Supplementary data is available at *ICES Journal of Marine Science* online.

Conflict of 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.

Funding

This work was supported by the Research Council of Norway under the Centre for Research-Based Innovation in Ma-

rine Acoustic Abundance Estimation and Backscatter Classification (CRIMAC) project (no. 309512).

Data availability

Data available on request: the data underlying this article are available on an S3 server and will be shared on reasonable request to the corresponding author. The code used for the different steps are stored on git and are listed in the supplementary information.

References

Brautaset O, Waldeland AU, Johnsen E *et al.*. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. *ICES J Mar Sci* 2020;**77**:1391–400. https://doi.org/10.1093/icesjms/fsz235

Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal* 2021;**71**:102062. https://doi.org/10.1016/j.media.2021.102062

Choi C, Kampffmeyer M, Handegard NO *et al.*. Deep semisupervised semantic segmentation in multifrequency echosounder data. *IEEE J Oceanic Eng* 2023;**48**:384–400.

Choi C, Kampffmeyer M, Handegard NO *et al.* Semi-supervised target classification in multi-frequency echosounder data. *ICES J Mar Sci* 2021;**78**:2615–27. https://doi.org/10.1093/icesjms/fsab140

De Robertis A, Higginbottom I. A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. *ICES J Mar Sci* 2007;**64**:1282–91. https://doi.org/10.1093/icesjms/fsm112

Demer DA, Berger L, Bernasconi M *et al.*. Calibration of acoustic instruments. *ICES Coop Res Rep* 2015;**326**:136.

Dunford AJ. Correcting echo-integration data for transducer motion. *J Acoust Soc Am* 2005;**118**:2121–3. https://doi.org/10.1121/1.2005927

Freeman S, Mackinson S, Flatt R. Diel patterns in the habitat utilisation of sandeels revealed using integrated acoustic surveys. *J Exp Mar Biol Ecol* 2004;**305**:141–54. https://doi.org/10.1016/j.jembe.2003.12.016

Furness RW. Management implications of interactions between fisheries and sandeel-dependent seabirds and seals in the North Sea. *ICES J Mar Sci* 2002;**59**:261–9. https://doi.org/10.1006/jmsc.2001.1155

Gunderson DR. *Surveys of Fisheries Resources*. New York, NY: John Wiley & Sons, 1993, 278pp.

Haralabous J, Georgakarakos S. Artificial neural networks as a tool for species identification of fish schools. *ICES J Mar Sci* 1996;**53**:173–80. https://doi.org/10.1006/jmsc.1996.0019

Johnsen E, Harbitz A. Small-scale spatial structuring of burrowed sandeels and the catching properties of the dredge. *ICES J Mar Sci* 2013;**70**:379–86. https://doi.org/10.1093/icesjms/fss202

Johnsen E, Kvamme C. Foreløpig råd for tobisfiske i norsk økonomisk sone i 2024–Faglig grunnlag. 47. *Havforskningsinstituttet*. 2024. https://imr.brage.unit.no/imr-xmlui/handle/11250/3126610(27 November 2024, date last accessed).

Johnsen E, Pedersen R, Ona E. Size-dependent frequency response of sandeel schools. *ICES J Mar Sci* 2009;**66**:1100–5. https://doi.org/10.1093/icesjms/fsp091

Johnsen E, Rieucau G, Ona E *et al.*. Collective structures anchor massive schools of lesser sandeel to the seabed, increasing vulnerability to fishery. *Mar Ecol Prog Ser* 2017;**573**:229–36. https://doi.org/10.3354/meps12156

Johnsen E, Totland A, Skålevik Å *et al.*. StoX: an open source software for marine survey analyses. *Methods Ecol Evol* 2019;**10**:1523–8. https://doi.org/10.1111/2041-210X.13250

Kloser RJ, Ryan T, Sakov P *et al.*. Species identification in deep water using multiple acoustic frequencies. *Can J Fish Aquat Sci* 2002;**59**:1065–77. https://doi.org/10.1139/f02-076

Komiyama S, Holmin AJ, Pedersen G *et al.*. Silent uncrewed surface vehicles reveal the diurnal vertical distribution of lesser sandeel. *ICES J Mar Sci* 2024;**82**:fsae159. https://doi.org/10.1093/icesjms/fsae159

R. Korneliussen (ed). Acoustic target classification. *ICES Coop Res Rep* 2018;**344**:104.

Korneliussen RJ, Heggelund Y, Eliassen IK *et al.* Acoustic species identification of schooling fish. *ICES J Mar Sci* 2009;**66**:1111–8. https://doi.org/10.1093/icesjms/fsp119

Korneliussen RJ, Heggelund Y, Macaulay GJ *et al.*. Acoustic identification of marine species using a feature library. *Methods Oceanogr* 2016;**17**:187–205. https://doi.org/10.1016/j.mio.2016.09.002

Korneliussen RJ, Ona E. Synthetic echograms generated from the relative frequency response. *ICES J Mar Sci* 2003;**60**:636–40. https://doi.org/10.1016/S1054-3139(03)00035-3

Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: F Pereira, CJC Burges, L Bottou, KQ Weinberger (eds), *Advances in Neural Information Processing Systems*, Vol. **25**. San diego, CA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2012, 1097–105.

Lecun Y, Bottou L, Bengio Y *et al.*. Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**:2278–324. https://doi.org/10.1109/5.726791

Macer CT. Sand eels (*Ammodytidae*) in the southwestern North Sea; their biology and fishery. Fishery Investigations Series 2, Vol. **24**. *Ministry of Agriculture, Fisheries and Food*, London. 1966; 55pp.

MacLennan D, Simmonds EJ. *Fisheries Acoustics*. *Fish and Aquatic Resources Series 10*. London: Chapman & Hall, 2005.

MacLennan DN, Holliday DV. Fisheries and plankton acoustics: past, present, and future. *ICES J Mar Sci* 1996;**53**:513–6. https://doi.org/10.1006/jmsc.1996.0074

Marques TP, Cote M, Rezvanifar A *et al.*. Instance segmentation-based identification of pelagic species in acoustic backscatter data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4378–87. IEEE, Virtual meeting, 2021a.

Marques TP, Rezvanifar A, Cote M *et al.*. Detecting marine species in echograms via traditional, hybrid, and deep learning frameworks. In: *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5928–35. Milan, Italy: IEEE, 2021b.

Nguyen TX, Winger PD, Orr D *et al.*. Computer simulation and flume tank testing of scale engineering models: how well do these techniques predict full-scale at-sea performance of bottom trawls? *Fish Res* 2015;**161**:217–25. https://doi.org/10.1016/j.fishres.2014.08.007

Ordoñez A, Utseth I, Brautaset O *et al.*. Evaluation of echosounder data preparation strategies for modern machine learning models. *Fish Res* 2022;**254**:106411. https://doi.org/10.1016/j.fishres.2022.106411

Pala A, Oleynik A, Utseth I *et al.*. Addressing class imbalance in deep learning for acoustic target classification. *ICES J Mar Sci* 2023;**80**:2530–44. https://doi.org/10.1093/icesjms/fsad165

Pala A, Oleynik A, Malde K, *et al.* Self-supervised feature learning for acoustic data analysis. *Ecological Informatics*, 2024;**84**:102878–. https://doi.org/10.1016/j.ecoinf.2024.102878

Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–71. IEEE, Honolulu, USA, 2017.

Reid D, Scalabrin C, Petitgas P *et al.*. Standard protocols for the analysis of school based data from echo sounder surveys. *Fish Res* 2000;**47**:125–36. https://doi.org/10.1016/S0165-7836(00)00164-8

Reid DG. *Report on Echo Trace Classification*. *ICES Cooperative Research Report No. 238*. International Council for the Exploration of the Sea, Copenhagen, Denmark, 2000.

Ren S, He K, Girshick R *et al.*. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*, p. 9. Curran Associates, Inc., Montréal, Canada, 2015. https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf(3 April 2025, date last accessed).

Rezvanifar A, Marques TP, Cote M *et al.*. 2019. A deep learning-based framework for the detection of schools of herring in echograms. arXiv:1910.08215[cs, eess, stat]. http://arxiv.org/abs/1910.08215(6 August 2021, date last accessed).

Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: N Navab, J Hornegger, WM Wells, AF Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Cham: Springer International Publishing, 2015, 234–41.

Strindberg S, Buckland ST. Zigzag survey designs in line transect sampling. *J Agric Biol Environ Stat* 2004;**9**:443. https://doi.org/10.1198/108571104X15601

Vohra R, Senjaliya F, Cote M *et al.*. Detecting underwater discrete scatterers in echograms with deep learning-based semantic segmentation. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 375–84. Vancouver, Canada, 2023.

Winslade P. Behavioural studies on the lesser sandeel *Ammodytes marinus* (Raitt) II. The effect of light intensity on activity. *J Fish Biol* 1974;**6**:577–86. https://doi.org/10.1111/j.1095-8649.1974.tb05101.x

Wright PJ, Jensen H, Tuck I. The influence of sediment type on the distribution of the lesser sandeel, *Ammodytes marinus*. *J Sea Res* 2000;**44**:243–56. https://doi.org/10.1016/S1385-1101(00)00050-2

Handling Editor: Pavanee Annasawmy