ICES Journal of Marine Science, 2023, **00**, 1–15
DOI: 10.1093/icesjms/fsad165
Original Article

Addressing class imbalance in deep learning for acoustic target classification

Ahmet Pala [1,*], Anna Oleynik [1], Ingrid Utseth [2], and Nils Olav Handegard [3]

[1]Department of Mathematics, University of Bergen, Allegaten 41, Bergen 5008, Norway
[2]Norwegian Computing Center, P. O. box 114 Blindern, Oslo 0314, Norway
[3]Institute of Marine Research, Nordnesgaten 50, Bergen 5005, Norway

*Corresponding author:Tel: +47 96745817; e-mail: ahmet.pala@uib.no.

Acoustic surveys provide important data for fisheries management. During the surveys, ship-mounted echo sounders send acoustic signals into the water and measure the strength of the reflection, so-called backscatter. Acoustic target classification (ATC) aims to identify backscatter signals by categorizing them into specific groups, e.g. sandeel, mackerel, and background (as bottom and plankton). Convolutional neural networks typically perform well for ATC but fail in cases where the background class is similar to the foreground class. In this study, we discuss how to address the challenge of class imbalance in the sampling of training and validation data for deep convolutional neural networks. The proposed strategy seeks to equally sample areas containing all different classes while prioritizing background data that have similar characteristics to the foreground class. We investigate the performance of the proposed sampling methodology for ATC using a previously published deep convolutional neural network architecture on sandeel data. Our results demonstrate that utilizing this approach enables accurate target classification even when dealing with imbalanced data. This is particularly relevant for pixel-wise semantic segmentation tasks conducted on extensive datasets. The proposed methodology utilizes state-of-the-art deep learning techniques and ensures a systematic approach to data balancing, avoiding ad hoc methods.

Keywords: acoustic target classification, big data, class imbalance, deep learning, similarity-based sampling.

Introduction

Vertically oriented sonars, also known as echo sounders (Korneliussen, 2018), are employed in acoustic trawl surveys to provide indices of abundance to fisheries assessment models (Simmonds and MacLennan, 2008). Typically, several transducers are operating simultaneously, where each transducer operates at a specific frequency or frequency range. The strength of the reflected signals from objects in the water column is commonly referred to as backscatter and is typically represented as single backscattering cross-section coefficients (MacLennan *et al.*, 2002). The collected backscatter data from each transducer are displayed in an echogram as a function of ping time and range (depth). Echograms provide a high-resolution depiction of the ocean interior, where underwater structures like the seabed, fish schools, zooplankton layers, and various other objects can be seen (Blackwell *et al.*, 2020). When the backscatter from a single fish species can be isolated, the backscatter is linearly related to fish abundance (Foote, 1983).

Despite the high spatial resolution of the echosounder, the taxonomic resolution is limited. Reliably allocating backscatter to an acoustic category, where the category may represent a species or species group, can be challenging (Korneliussen, 2018). Traditionally, this has been a manual process based on experience, often with the aid of trawl sampling (Simmonds and MacLennan, 2008). Successful examples include sandeel (Johnsen *et al.*, 2009), herring (Karp and Walters, 1994), and blue whiting (Gastauer *et al.*, 2016), among others. Although this method has proven successful in many cases, it is time-consuming and can introduce biases due to the operator's subjective judgement or personal preferences.

To avoid bias and subjectivity, automated methods have been suggested. These methods utilize algorithms to automatically classify targets (c.f. Korneliussen, 2018) for a recent review. This process is coined Acoustic Target Classification (ATC) and includes methods that rely on the frequency response of targets (Kloser *et al.*, 2002; Korneliussen, 2002) or shape, depth, and other features derived from the acoustic backscatter signals (Haralabous and Georgakarakos, 1996; Reid, 2000). The recent developments in machine learning methods, including deep learning methods (LeCun *et al.*, 2015), have started to gain momentum for ATC.

Several machine learning methods have been proposed for ATC. A framework proposed by Rezvanifar *et al.* (2019) introduces a region of interest extractor combined with a deep learning-based image classifier. Others have used similar approaches where the region of interest is first detected, followed by ATC for the detected region. As Marques *et al.* (2021b) proposed, the steps can be combined using end-to-end deep learning frameworks like Faster R-CNN (Ren *et al.*, 2015) and YOLOv2 (Redmon and Farhadi, 2017). Additionally, YOLOv8 (Talaat and ZainEldin, 2023) is a more recent version that is accessible for the particular task of object detection. Training these networks requires a substantial amount of labelled data, and Choi *et al.* (2021) proposed a semi-supervised deep learning method to reduce the amount of data required. Marques *et al.* (2021a) proposed a deep learning-based instance segmentation framework to accurately identify bounding boxes for herring schools. Another class of models

Received: 13 July 2023; **Revised:** 28 September 2023; **Accepted:** 3 October 2023

provides semantic segmentation, where the acoustic category is predicted for each sample (pixel) in the echogram. Brautaset *et al.* (2020) developed a method based on the U-Net algorithm, and Ordonez *et al.* (2022) tested the performance of this model by exploring different resolutions and incorporating depth information. Vohra *et al.* (2023) tested Attention U-Net, U-Net, and DeepLabV3 for semantic segmentation, and Choi *et al.* (2023) combined semantic segmentation with unsupervised learning. It is worth emphasizing that the quality of the training data significantly affects the performance of the model. Insufficient or unrepresentative data leads to poor generalization, while richer, more representative data increases performance even with simpler algorithms (Mumuni and Mumuni, 2022). Obtaining such data is challenging, and data augmentation techniques that provide additional training samples by, for example, affine transformations such as scale and rotation (Wong *et al.*, 2016), introducing noise (Summers and Dinneen, 2019), or even using generative models (Wang *et al.*, 2021), have emerged as a solution to increase model robustness and accuracy. It is important to note that some augmentation techniques, like rotation, typically used for images are not appropriate for acoustic data.

A challenge for deep learning for ATC, and particularly for semantic segmentation methods, is class imbalance. Class imbalance occurs when one class is highly underrepresented, which is the case in ATC, where the target species are few and far between compared to the empty water column, seabed, and other structures, collectively referred to as background. For an overview of problems and methods related to the class imbalance, we refer the reader to Krawczyk (2016) and Buda *et al.* (2018). Approaches to address the class imbalance include resampling techniques, i.e. data-level approaches (Beyan and Fisher, 2015), cost-sensitive learning, and algorithmic approaches (Hasib *et al.*, 2020). Cost-sensitive learning involves assigning different misclassification costs to different classes (Zhou and Liu, 2010). Algorithmic approaches involve modifying the learning algorithm to handle class imbalance directly (Krawczyk, 2016). Resampling techniques work by resampling the unbalanced training dataset before the model training phase. The original unbalanced dataset can be balanced by either oversampling (Chawla *et al.*, 2002; Hu *et al.*, 2009) the under-represented class, referred to as the minority class, or undersampling (Japkowicz, 2000) the over-represented class, referred as the majority class. In their most basic versions, random oversampling duplicates random samples from the minority class while random undersampling removes random samples from the majority class (Van Hulse *et al.*, 2007). Random undersampling has been shown to be preferable to random oversampling (Galar *et al.*, 2011; Błaszczyński and Stefanowski, 2015). This is due to the possibility of overfitting during the model generation process being increased by the oversampling method (Lin *et al.*, 2017).

Although random undersampling is widely used for imbalanced datasets, the disadvantage is that it could leave out data samples that are informative (Ng *et al.*, 2014). Several strategies to get around these limitations were proposed. Tomek (1976) proposed an undersampling method that removes the majority samples in close proximity to examples from the minority class, known as Tomek links, to improve the decision boundary of a classifier. Kubat (1997) proposed one-sided selection as a method to eliminate noisy and redundant samples from the majority class using a 1-nearest neighbours (1-NN) rule. Barandela *et al.* (2004) proposed a K-NN rule for eliminating misclassified samples from the training set, with a specific emphasis on removing the majority samples located at class boundaries. In order to gather distributional data and improve resampling diversity, Ng *et al.* (2014) devised a diversified sensitivity-based undersampling method. Sowah *et al.* (2016) proposed a cluster-based undersampling approach that removes repeated and noisy instances as well as outliers from the majority class. NearMiss is another undersampling technique that selects majority class samples closest to the minority class using a K-NN approach (Mani and Zhang, 2003). It promotes reliable and equitable class decision boundary (Bao *et al.*, 2016; Peng *et al.*, 2019) and has been successfully used in applications such as electricity theft detection in smart grids (Ullah *et al.*, 2022) and chronic kidney disease detection (Salau *et al.*, 2023).

Different approaches have been used to address data imbalance in machine learning methods for ATC. Brautaset *et al.* (2020) tried various approaches, including loss weighted by backscatter, with limited success. They resorted to an ad hoc sampling methodology where they randomly selected the background class with a higher probability close to the bottom. However, they still had problems with high-backscattering-intensity areas close to the surface being allocated to sandeel. Misclassifying the bottom is a significant error in ATC, but it is detectable by examining spikes in the summed predictions over range as a function of ping time. In contrast, identifying scattering that resembles the foreground class is far more challenging. Choi *et al.* (2021) addressed the class imbalance problem by randomly undersampling the background class and Vohra *et al.* (2023) employed Dice Loss with Binary Cross Entropy for the U-Net and the Attention U-Net algorithm and Focal Loss for DeepLabV3+. Choi *et al.* (2023) introduced a class-rebalancing weight for each learning objective.

The main objective of this paper is to address the issue of class imbalance for semantic segmentation networks in ATC by introducing an adaptable sampling approach. The proposed strategy aims at prioritizing background regions that have similar backscattering intensity characteristics as the foreground class. Through this approach, we expect to improve the training and achieve more accurate target classification when the datasets are imbalanced. To evaluate our proposed sampling methodology, we test it on the deep convolutional neural network proposed by Brautaset *et al.* (2020) and compare the performance with their sampling approach.

Material and methods

Acoustic data

In this paper, we use the acoustic data obtained from acoustic trawl surveys for sandeel (*Ammodytes marinus*). Sandeel is a swimbladder-less small fish that plays an important role in the North Sea ecosystem by serving as a primary food source for various predators, such as seabirds, seals, and larger fish (Furness, 2002). It also holds significant economic value as a target species for commercial fishing. The sandeel data used in this study were collected in the northeastern part of the North Sea by the Norwegian Institute of Marine Research (Johnsen *et al.*, 2017). The surveys were carried out using a Simrad EK60 with frequencies of 18, 38, 120, and 200 kHz, spanning the years 2007–2018. The surveys were operated with RV Johan Hjorth for 2007, 2008, 2010, and 2011, while RV GO

Sars and FV Brennholm were used for the surveys in 2009 and 2012, respectively. The survey series from 2013 to 2018 were conducted using FV Eros. Throughout the surveys, a 1.024 ms pulse duration, and a 3–4 Hz ping repetition frequency were used for all frequencies, and the vessel maintained a speed of approximately ten knots. Refer to Brautaset *et al.* (2020) for further details.

The echograms for each year were stored separately, and consist of the volume backscattering coefficients, s_v, arranged by ping time, range (depth), and frequency.

The backscattering coefficient is the average of backscattering intensity per cubic metre, i.e.

$$s_v = \frac{\sum \sigma_{bs}}{V}, \tag{1}$$

where σ_{bs} is the backscattering cross section, in m^2, while V corresponds to the volume occupied by a scattering medium, in m^3, see for details MacLennan *et al.* (2002). As a common data preprocessing step, see Lurton (2002) and MacLennan *et al.* (2002), the s_v values are transformed into volume backscattering strength values, S_v, also known as backscattering intensity, measured in decibels ($dB\ re\ 1m^{-1}$). The thresholded S_v values are calculated as

$$S_v = \begin{cases} 0 & \text{if} \quad 10\log_{10}(s_v) > 0 \\ -75 & \text{if} \quad 10\log_{10}(s_v) < -75 \\ 10\log_{10}(s_v) & \text{otherwise.} \end{cases} \tag{2}$$

We refer to a location on the echogram, in the ping time and range axes, as a pixel. Thus, each pixel has four S_v values that correspond to the four frequency channels.

During the data preprocessing, we standardized the data to be comparable between the years. Missing pings were identified using the median ping rate, and when found, columns of zeros were inserted into the s_v data. When the range resolution of other frequencies was lower, the data were interpolated onto the 200-kHz range vector. Conversely, if the range resolution was higher, s_v values were averaged into bins defined by the 200-kHz range vector. In cases where the pulse duration deviated from the standard settings, the data were interpolated onto the range grid for the standard setting. This process resulted in s_v values organized into a uniform time-range grid, similar to pixels in a four-channel image. The seabed location was approximated as the depth with the maximum increase in vertical gradient for each ping. Refer to Brautaset *et al.* (2020) for further details.

The echograms were manually labelled using the Large Scale Survey System (LSSS) (Korneliussen *et al.*, 2016). Based on these labels, the pixels are assigned to three distinct classes: the target species class as a "sandeel", an "other" class comprising all other fish species, and a "background" class encompassing all other objects visible in the echogram, such as the seabed and zooplankton layers.

The annotated echograms exhibit a strong imbalance in class distributions, with an overwhelming majority of pixels (99.8%) belonging to the "background" class, while a small fraction of pixels is annotated as either "sandeel" (0.1%) or "other" (0.1%). This dataset corresponds to the dataset used in Brautaset *et al.* (2020).

For this study, the data were split into training, validation, and test sets. We used separate years as training, validation, and test data. In particular, the survey data from 2011 to 2016 were used as the training data, 2017 for validation, and 2007, 2008, 2009, 2010, and 2018 for testing.

Neural network

We use a deep convolutional neural network that was designed for ATC on sandeel (Brautaset *et al.*, 2020). This network is a modified version of the U-Net, which was introduced for pixel-wise image segmentation tasks (Ronneberger *et al.*, 2015). For ATC, however, it is not feasible to train the network using the entire dataset as one "image" due to the extensive size of acoustic data. Therefore, fixed-size patches were selected from the annotated echograms. The patch size was set to 256 × 256 pixels.

The architecture of the network consists of an encoder and a decoder (c.f. Figure 1). First, it encodes the input into a 16 × 16 image with 1024 abstract features (1024 × 16 × 16). Then the decoder generates an output, classifying each input pixel into one of three categories: "background", "other", or "sandeel" (3 × 256 × 256). This output models the softmax probability for each pixel to belong to the three classes.

Depending on how the training and validation patches are sampled, the resulting predictive model will be different. Below, we review the sampling procedure from Brautaset *et al.* (2020). We refer to this sampling approach as baseline sampling and call the corresponding trained U-Net as the baseline model (see Section 3.3). In Sections 3.4–3.5, we define three other sampling approaches and the corresponding predictive models: the similarity-based, random, and mixed model.

Baseline model

To address the class imbalance problem, Brautaset *et al.* (2020) proposed to sample input patches according to six different patch classes (c.f. Table 1). A patch class is a label that is associated with each patch for the training step. They used equal sampling probabilities for all the patch classes except the "Background" class, the class of patches composed of only "background" pixels. "Background" patches were selected at random but ensuring that no fish schools or seabed were present within the patch. For the "Seabed" class, random patches containing seabed were chosen. Both the "Sandeel" and "Other" classes involved selecting a random pixel within a fish school to draw the patch around it in the way that the patch would typically cover the fish school. For the other two classes, "Seabed + Sandeel" and "Seabed + Other", the same procedure was used except that only the fish schools close to the seabed were chosen. This ensured that the patches included the seabed.

In this undersampling approach, only the background patches containing seabed were considered important. Thus, the other diverse "background" structures, such as zooplankton layers were not explicitly taken into account. This leads the baseline model to commonly misclassify the zooplankton layers as fish schools. Since the non-fish school objects are not annotated, improving this baseline sampling is not realistic. In addition, determining the optimal probability values for each proposed patch class is not straightforward. Hence, another approach that can pinpoint significant regions for extracting patches in the sampling stage is needed.

Data-driven similarity-based sampling approach

Our approach is to improve the sampling of the "background" by prioritizing the regions that have similar characteristics to

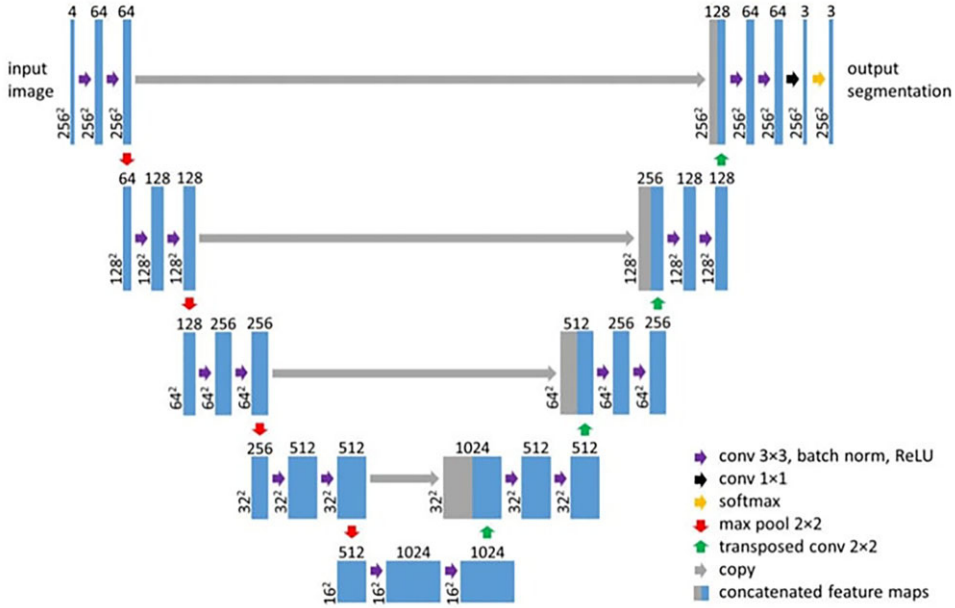


Figure 1. The structure of the model used in this study (Brautaset *et al.*, 2020). The network is designed to process input data consisting of four channels representing frequencies of 18, 38, 120, and 200 kHz. Each channel has dimensions of 256 × 256. The network's output is a 3 × 256 × 256 tensor, where each pixel is assigned softmax probabilities indicating its belonging to one of the classes: "sandeel", "other", or "background".

Table 1. Proposed patch classes together with definitions and sampling probabilities in Brautaset *et al.* (2020).

Patch classes	Probability	Description
Background	1/26	Random patch from area without fish, above the seabed
Seabed	5/26	Random patch from area containing seabed
Sandeel	5/26	Random patch from area containing "sandeel" class
Other	5/26	Random patch from area containing "other" class
Seabed + sandeel	5/26	Random patch from area containing both "seabed" and "sandeel" classes
Seabed + other	5/26	Random patch from area containing both "seabed" and "other" classes

The regions of the echograms were classified into six distinct classes, and random samples were drawn from each class according to the provided probabilities.

"sandeel" schools in terms of backscattering intensity. Instead of using manually designed patch classes, we use the classes corresponding to the pixel classes, namely, patches containing "sandeel", "other", and "background" pixels. The patch classes are sampled from the echograms with equal probability (see Table 2).

It is rather straightforward to sample patches containing "sandeel" and "other" since these classes are annotated. Sampling patches targeting the "background", which include various unannotated structures with high variation in the backscattering intensities is however challenging. Random sampling does not guarantee that these structures will be included nor to depict the variation in the backscattering intensities. For example, if the echogram range is much larger than the patch height, the majority of the randomly selected patches would represent only empty water and have low backscattering intensity values.

We propose to sample background regions that have similar backscatter properties as the "sandeel" and, thus, are prone to misclassification. In order to accomplish this, we employed the NearMiss undersampling methodology (Mani and Zhang, 2003). This methodology comprises three versions, among which NearMiss-1 selects samples from the majority class that have the smallest average distance to the three closest samples from the minority class. This approach helps to remove majority-class samples that are most likely to be misclassified as minority-class samples. NearMiss-2 selects samples from the majority class that have the smallest average distance to the three furthest samples from the minority class. This approach helps to preserve more information from the majority class while still reducing the imbalance ratio. NearMiss-3 selects a given number of majority-class samples for each example in the minority class that are closest. This approach is particularly useful when the minority class samples are scattered throughout the feature space (Mani and Zhang, 2003). Here we focus on NearMiss-2 as it is less biassed towards the distributions within the majority or minority class.

The majority and minority classes and their samples must be properly defined. Since sandeel is our target species, the "sandeel" class is considered as the minority class while the "background" represents the majority class. The "other" class could be viewed as a part of the majority class. However, since it is small in size and annotated, we excluded it from the consideration. If we considered "other" as a part of the majority class, then applying undersampling on "other" might exclude fish species that are similar to sandeel just because they are less similar than some pixels in the background class, e.g. zooplankton layers. Thus, the other fish schools would not be well represented. If we applied undersampling on "other"

Table 2. The proposed similarity-based sampling strategy.

Patch classes	Probability	Description
Similarity-based background	1/3	Random patch centred by the NearMiss regions
Sandeel	1/3	Random patch from area containing "sandeel" class
Other	1/3	Random patch from area containing "other" class

Three patch classes are proposed, corresponding to the pixel classes. "Similarity-based Background" patches contain regions similar to sandeel fish schools. The "Sandeel" and "Other" patch classes contain fish schools of their respective classes.

Table 3. Number of majority (background regions, $\mathbf{y}_j$) and minority samples (sandeel schools, $\mathbf{x}_t$) for training and validation years.

Survey year	Majority ($\mathbf{y}_j$)	Minority ($\mathbf{x}_t$)
2011	938,682	623
2013	860,413	2,015
2014	1,369,136	1,121
2015	67,6913	1,515
2016	1,412,651	829
2017	743,134	3,602

separately, it would eliminate the useful samples as the class is small enough.

Regarding the construction of minority class samples, we averaged the backscattering intensity for each frequency over each annotated "sandeel" school, i.e.

$$\mathbf{x}_t = \frac{\sum_{i \in I_t} \mathbf{S}_{\mathbf{v}_i}}{\mid I_t \mid}, \quad t = 1, 2, \ldots, T, \tag{3}$$

where $\mathbf{S}_{\mathbf{v}_i}$ is a four-dimensional vector of S_v values of the pixel i, I_t represents the index set for a fish school t, and T is the total number of sandeel fish schools. The averaging ensures that any random fluctuations are removed, similar to the approach that Korneliussen (2000) used. Note that we average S_v values since they, and not the s_v, are the inputs to the neural network. The number of sandeel fish schools T varies between 623 and 3602 over the training and validation years (c.f. Table 3).

To accurately define the majority class, we excluded all annotated fish schools and pixels below the seabed. Using individual pixel values for the majority class would yet again be prone to random fluctuations in the data. Therefore, we averaged the "background" S_v values over a specific window size. If the window size is too small, the resulting values may still be too variable as well as computationally expensive. Conversely, if we opt for a window size that is excessively large, the resulting values may not be sufficiently representative of the "background" S_v value distribution. Given that the choice of window size is a trade-off between computational efficiency and accuracy, and considering the average size of sandeel fish schools, we use a window of 25×25 after experimenting with different window sizes. Then, we computed the mean values of the 25×25 pixel regions for each frequency. Thus, we obtained four-dimensional vectors $\mathbf{y}_j, j = 1,\ldots, J$, representing the majority class samples. Here J is the number of non-overlapping 25×25 regions covering the echogram dimensions. The procedure is depicted in Figure 2 for one frequency channel.

For each acoustic dataset in the training and validation sets, once we have defined the majority and minority classes, we calculated ρ_t^j values representing the Euclidean distances from each majority sample $\mathbf{y}_j$ to every individual minority sample $\mathbf{x}_t$, i.e. $\rho_t^j = \|\mathbf{y}_j - \mathbf{x}_t\|$, $j = 1, \ldots, J$, and $t = 1, \ldots T$. We then defined the ordered set $D_j = \{\rho_1^j, \rho_2^j, \ldots, \rho_T^j\}$, for each of the majority samples $\mathbf{y}_j, j = 1, \ldots, J$, containing ρ_t^j values in

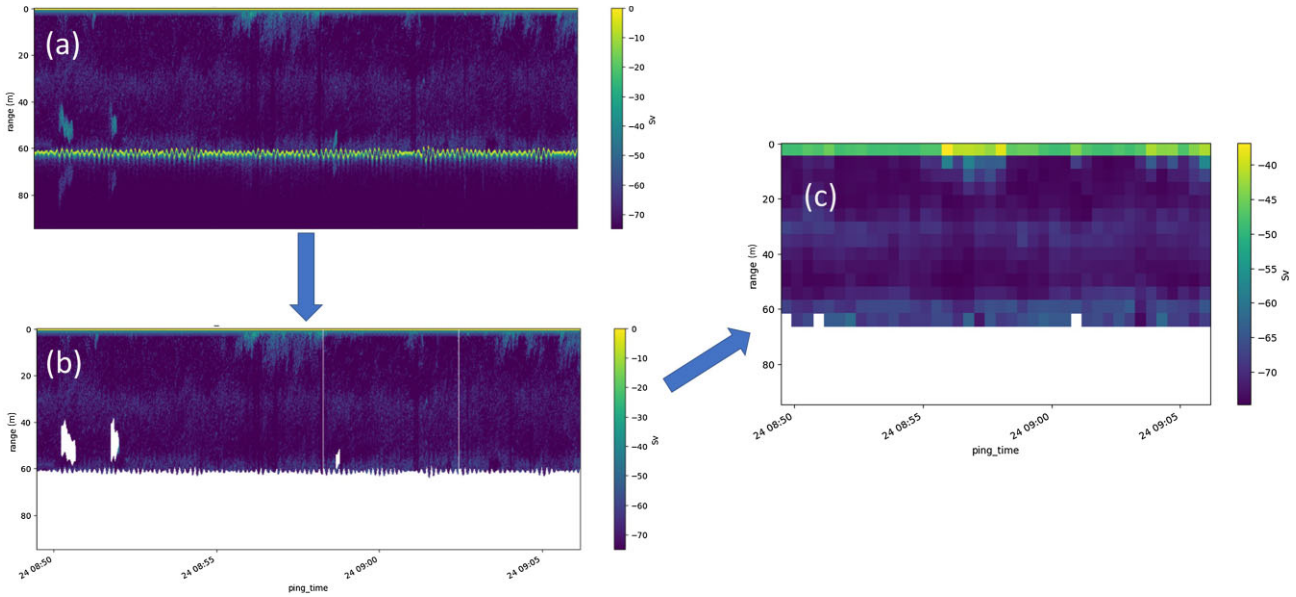


Figure 2. The steps of averaging for "background" pixels are demonstrated on the example echogram at 200 kHz frequency, denoted as (a). In this process, the fish school pixels and the pixels below the seabed are excluded from the echogram (b). Subsequently, the remaining pixels are averaged using 25×25 pixel window (c).

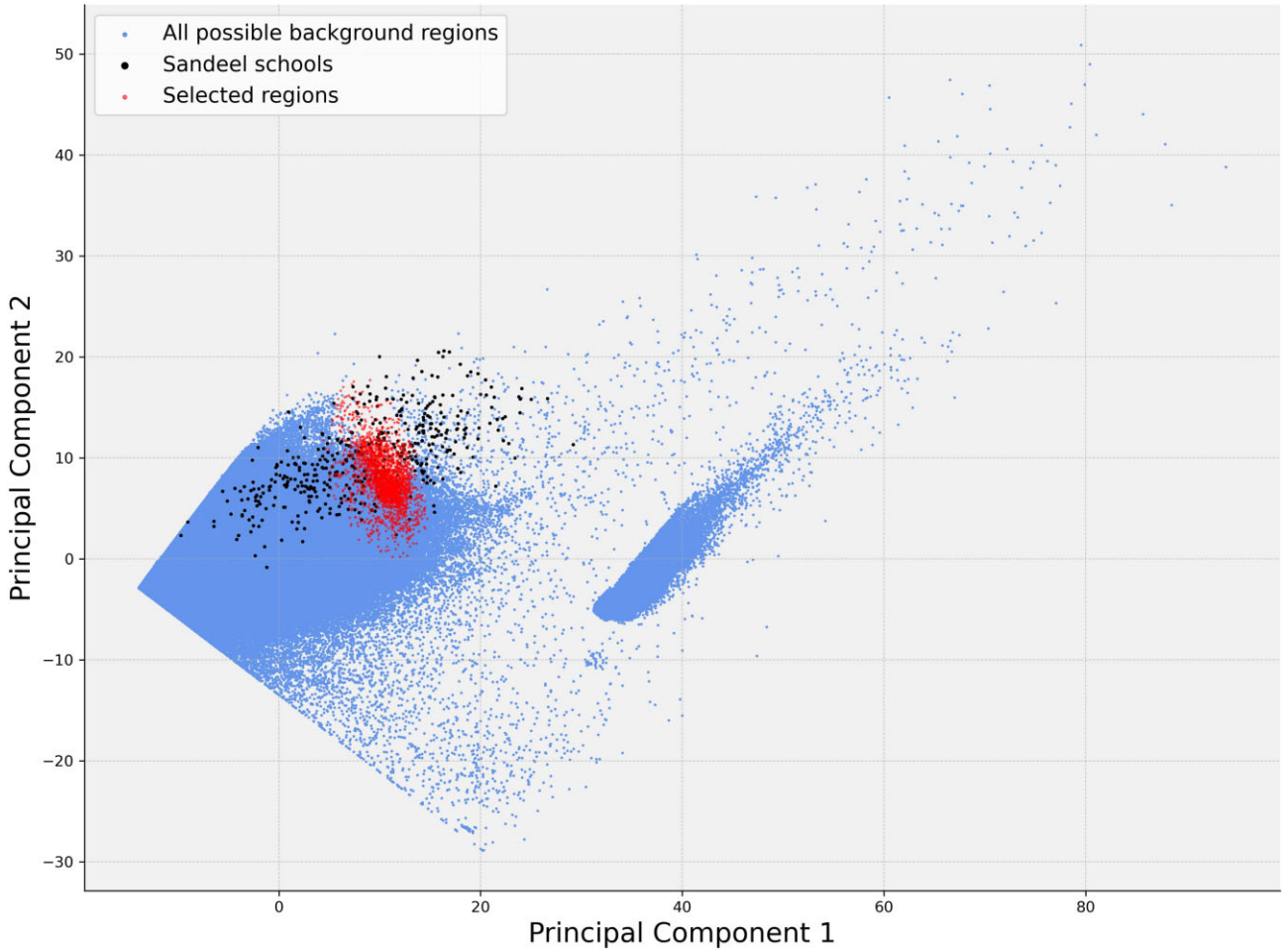


Figure 3. The scatter plot illustrates the results of PCA after applying the NearMiss-2 algorithm to the acoustic data for a single survey. The majority class ($\mathbf{x}_j$) is visualized as blue points on the two principal components. The minority class ($\mathbf{x_t}$), which represents data points related to sandeel fish schools, is depicted in black. Additionally, the selected 25 × 25 background regions from I_{SB} are shown as red points (highlighted by red squares in Figure 4).

descending order, where $\rho_k^j \leq \rho_{k+1}^j$, $k = 1, \ldots, T\text{-}1$. From this distance set, we calculate the average, $\boldsymbol{\mu}_j$, of the first $S = 3$ values, i.e.

$$\mu_j = \frac{1}{S}\sum_{s=1}^{S} \rho_s^j, \quad \rho_s^j \in D_j, \quad j = 1, \ldots, J. \tag{4}$$

These values were then sorted in ascending order and the Nth smallest value was denoted as m_N. Our aim here is to incorporate the most similar regions while also introducing some level of variation. To achieve this, we conducted experiments with different values of N and determined that setting N at five times the number of fish schools in each training dataset gave the best F1 score on the validation data. Finally, the similarity-based index set I_{SB} was defined as

$$I_{SB} = \left\{ j \mid \mu_j \leq m_N \right\}. \tag{5}$$

This set points to the most similar, according to the NearMiss-2 methodology, background regions to the sandeel fish schools in the acoustic data.

To demonstrate the distribution of the selected $\mathbf{y}_j$, $j \in I_{SB}$, we performed principal component analysis (PCA) (Abdi and Williams, 2010) on the majority, $\mathbf{y}_j$, $j = 1, \ldots, J$, and the minority, $\mathbf{x}_t$, $t = 1, \ldots, T$, samples. The scatter plot in Figure 3 shows that the selected majority samples cluster around the centre of the minority class. This is expected since we have selected the minimum averaged distances of the three furthest distances to the minority class.

We also attempted NearMiss-1 and NearMiss-3 techniques to undersample the majority class. However, NearMiss-1 led to the selected majority samples resembling the distribution of the minority class. This outcome was not desirable as our objective was to capture the most typical sandeel fish schools, rather than extreme cases with very high or very low backscattering intensities. Similarly, NearMiss-3 followed the distribution of the majority class, but its emphasis on the "background" class could result in selecting irrelevant areas. On the other hand, NearMiss-2 allowed us to concentrate more on the most similar regions at the centre of the minority class. Additionally, as we show below, our sampling methodology incorporates the surrounding pixels, thereby already accounting for intensities variation.

The training step requires 256 × 256 patches. Therefore, we selected the 256 × 256 patches centred on the 25 × 25 regions corresponding to $\mathbf{y}_j$, where j represents randomly sampled indices from the set denoted as I_{SB}. We refer to this sampling procedure as the similarity-based background sampling and the corresponding patch class as the "Similarity-Based Background" patch class. We refer to the overall sampling approach (see Table 2), as the similarity-based sampling, and the

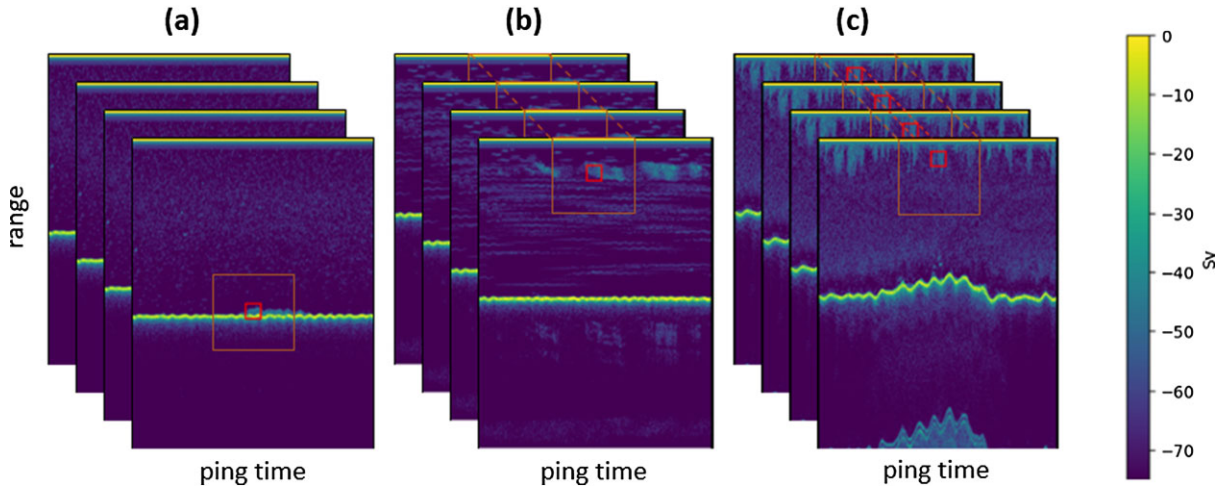


Figure 4. Three example patches (a, b, and c) where three "Similarity-Based Background" patches are generated based on the similarity-based sampling. The inner red squares are the 25 × 25 regions from I_{SB} that the NearMiss-2 algorithm determined similar to sandeel schools, while the outer orange regions are the 256 × 256 input patches used during training. Note that the background patches may include pixels of wider range of backscatter intensities, while the centre patch ensures that the difficult pixels are included during training.

corresponding network model as the similarity-based model. The "Sandeel" and "Other" patches are sampled as the baseline model, c.f. the description in Section 3.3. We emphasize that the exclusion of the aforementioned pixels is done only to determine the coordinates for the similarity-based sampling methodology. Once the coordinates on the acoustic data have been identified, patches are drawn from the whole echograms without any pixel exclusion.

The resulting "Similarity-Based Background" patch class includes patches with background regions having similar backscattering intensity patterns to those of sandeel fish schools. Figure 4 shows examples of the patches from the "Similarity-Based Background" patch class. There are cases where the annotations are missing, and Figure 4a provides an example where the selected patch contains an unannotated sandeel fish school. In Figure 4b and c, the identified regions are representing dense plankton layers.

When training the network with these "Similarity-Based Background" patches, we expect the resulting model to improve. Moreover, since a "Similarity-Based Background" patch also includes surrounding pixels (c.f. Figure 4), the network will also be exposed to morphological features that will provide additional information during training. In addition, these surrounding pixels exhibit the characteristic traits of the "background" class ensuring that the network is adequately exposed to the more common parts of the "background" class as well.

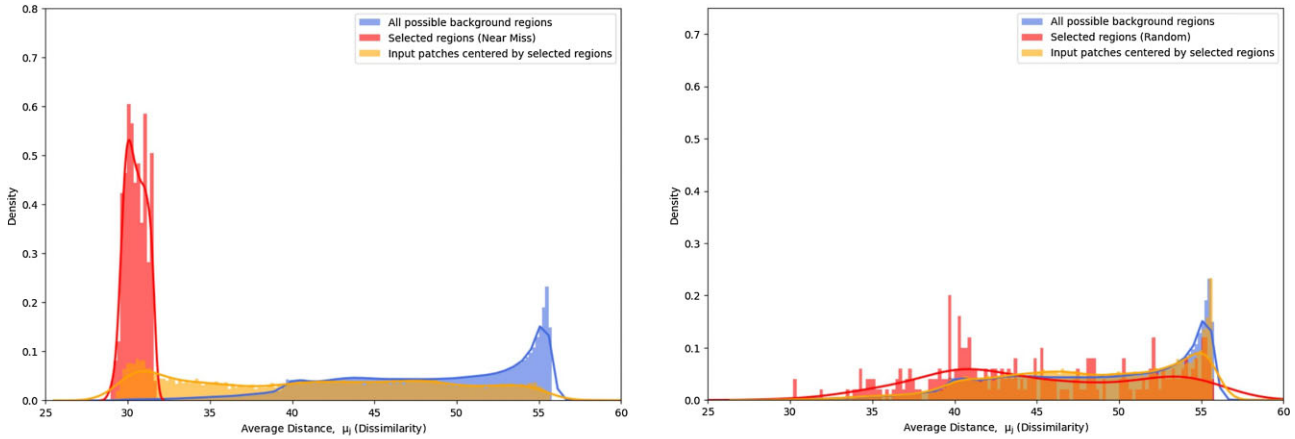
To assess the performance of the similarity-based undersampling method, we compared the distribution of μ_j based on the similarity-based sampling and that of random sampling (c.f. Figure 5) for one survey (2011). The figure maintains the same colour scheme as Figure 4, using red for the selected regions and orange for the input patches for easy identification and comparison. The histogram of the selected μ_j, $j \in I_{SB}$, which corresponds to the regions of interest, is depicted in red (see Figure 5a). These regions represent the most sandeel-like regions in the data, such as not-annotated sandeel fish schools (as given in Figure 4a) and zooplankton layers (as given in Figure 4b and c). In Figure 5b, the histogram in red corresponds to a random selection μ_j, and thus to random 25 × 25 background regions. By drawing 256 × 256-sized patches centred around the selected regions, we obtain the corresponding μ_j values for the regions within the patches. The histograms for these values are displayed in orange colour. The resulting distribution in Figure 5a is similar to a uniform distribution but with a slightly higher density of samples that resemble those of sandeel schools. The distribution resulting from the random selection (see Figure 5b) follows the majority class distribution. This indicates that random selection can capture some degree of diversity in the acoustic data, but the resulting regions used as inputs for the networks closely resemble the distribution of the majority class. Due to the low density of the majority class distribution for low μ_j, the most similar to the sandeel school regions can be missed.

Other sampling approaches

We tested a total of four different sampling strategies for training the neural network. In addition to the baseline sampling where patches that include the bottom were emphasized (c.f. Section 3.3), and the similarity-based sampling approach (c.f. Section 3.4), we tested a random sampling methodology for the "background" class. We also tested a combination of the random and similarity-based where we sampled the "Similarity-Based Background" patches and the random "Background" patches with a probability of 1/6 each. The corresponding trained models for these random and mixed sampling methodologies are referred as the random and mixed models, respectively (c.f. Table 4). We employed these models to assess our proposed sampling method against the widely used random undersampling technique and to study how the results change when the two sampling strategies are combined. Please note that all sampling methods are applied only to the training and validation datasets, and these datasets are chosen before undersampling is carried out.

Training and evaluation procedures

We trained the network using the data obtained by the four sampling strategies: baseline, similarity-based, random,



(a) "Similarity-Based Background" patch selection (b) Random "Background" patch selection

Figure 5. The distributions of $\boldsymbol{\mu}_j$ for "Similarity-Based Background" patch selection (a) and random "Background" patch selection (b). The distribution of $\boldsymbol{\mu}_j$ for all the majority samples, i.e. $j = 1,\ldots, J$, is represented by the blue colour. The red colour in (a) indicates $\boldsymbol{\mu}_j$, $j \in I_{SB}$, while the same colour in (b) represents the randomly selected $\boldsymbol{\mu}_j$. The orange colour corresponds to $\boldsymbol{\mu}_j$ of the regions contained in the selected 256 × 256 patches.

Table 4. The sampling strategies employed to obtain random 256 × 256 patches for training the similarity-based, random, and mixed models are described.

Patch classes	Probability			Description
	Similarity-based	Mixed	Random	
Sandeel	1/3	1/3	1/3	Random patch from area containing "sandeel" class
Other	1/3	1/3	1/3	Random patch from area containing "other" class
Similarity-based background	1/3	1/6	0	Random patch centred by the NearMiss regions
Background	0	1/6	1/3	Random patch from area without fish, above seabed

Additionally, the sampling probabilities for each patch class are provided, along with an explanation of the criteria used.

and mixed sampling. Each time, we apply the same training and validation configurations as Ordonez *et al.* (2022). In summary, we trained the network using stochastic gradient descent with a batch size of 16, an initial learning rate of 0.005 (reducing this value every 1000 iterations by a factor of 2) and a momentum value of 0.95. To further deal with the class imbalance in the dataset, the network employs a weighted cross-entropy loss function with class weights of "background", "sandeel", and "other", to be 10, 300, and 250, respectively to train the models on 160000 randomly chosen samples based on the sampling schemes. As a stopping criterion, the training iterations were limited to 10000.

All the trained models are binary pixel classifiers with a threshold between 0 and 1. Pixels are classified as positive if the output for the "sandeel" class exceeds the threshold and negative otherwise. To evaluate the performance of the models, we utilized precision–recall curves. Precision is the ratio of correctly predicted "sandeel" pixels among all predicted "sandeel" pixels, while recall is the proportion of accurately predicted "sandeel" pixels out of all the true "sandeel" pixels. By selecting the threshold that maximizes the F1 score, which balances precision and recall, we report the achieved F1 score as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (6)$$

To obtain classification results at the pixel level, we followed the same prediction implementation as Brautaset *et al.* (2020). This involved calculating precision and recall curves for each year separately. To generate the pixel-level classification results, we applied the trained models to small overlapping image patches. As a post-processing step, we eliminated any fish predictions located more than ten pixels below the seabed, i.e. classifying them as "background". The reason for this step is that pixels below the seabed represent the reflections of the objects located above the seabed.

We used the trained models, namely the baseline model and the new similarity-based, random, and mixed models, for producing predictions for all the acoustic data used for training, validation, and testing purposes. We obtained precision and recall curves by considering all pixels from the echograms for each year separately and we compared the performance of the models.

Results

In this section, we evaluate how the four sampling strategies for the training data influence the performance of the corresponding predictive models. Namely, the performance of our proposed similarity-based model was evaluated and compared to the baseline, random, and mixed models using maximized F1 scores across multiple survey years, as shown in Table 5. Precision–recall curves are given in Figure 6. We also examined at which echogram range (depth) the similarity-based model improved the "sandeel" pixel predictions compared to the baseline and random models in Figure 7. Additionally, we provide an example prediction from the baseline and our proposed model in Figure 8. It is important to note that, due to the undersampling, most of the background data has not been used for training and validation. The F1 scores for these years are a good indication of the model performances and

Table 5. F1 scores for each year are presented in bold numbers for the four aforementioned models.

Survey year	F1 scores			
	Baseline	Random	Mixed	Similarity-based
2007	0.1095	0.1655	0.3044	**0.4171**
2008	0.6622	0.6700	0.6629	**0.7073**
2009	0.8088	**0.8450**	0.8195	0.8137
2010	0.7704	**0.7777**	0.7728	0.7664
2011	0.7114	0.7784	0.7349	**0.7809**
2013	0.4776	0.5970	0.5848	**0.6347**
2014	0.7927	0.7978	0.7909	**0.8188**
2015	0.6320	0.6413	0.6323	**0.6530**
2016	0.4810	0.4799	0.4797	**0.5016**
2017	0.7495	0.8053	0.7789	**0.8164**
2018	0.8310	**0.8438**	0.8362	0.8311

The similarity-based model demonstrates better performance compared to the other models in most years, except for 2009, 2010, and 2018, where the performances are closely competitive.

therefore are reported together with the model performances on the test data.

The results demonstrate the effectiveness of the similarity-based sampling when compared to the baseline model, as it consistently achieves better performance across the survey series for training, validation, and test sets. Compared to the baseline model, the similarity-based model consistently achieves the best performance across the survey series except the test year of 2010. In particular, the similarity-based model exhibited considerable improvements for several years. For example, in the test year of 2007, the similarity-based model achieved an F1 score of 0.4171, which is ∼0.3076 higher than the baseline model's score of 0.1095. Similarly, in the training year of 2013, the model achieved an F1 score of 0.6347, representing an improvement of ∼0.1571 compared to the baseline model.

In terms of the F1 score, the random model performance was slightly better than that of the similarity-based model for

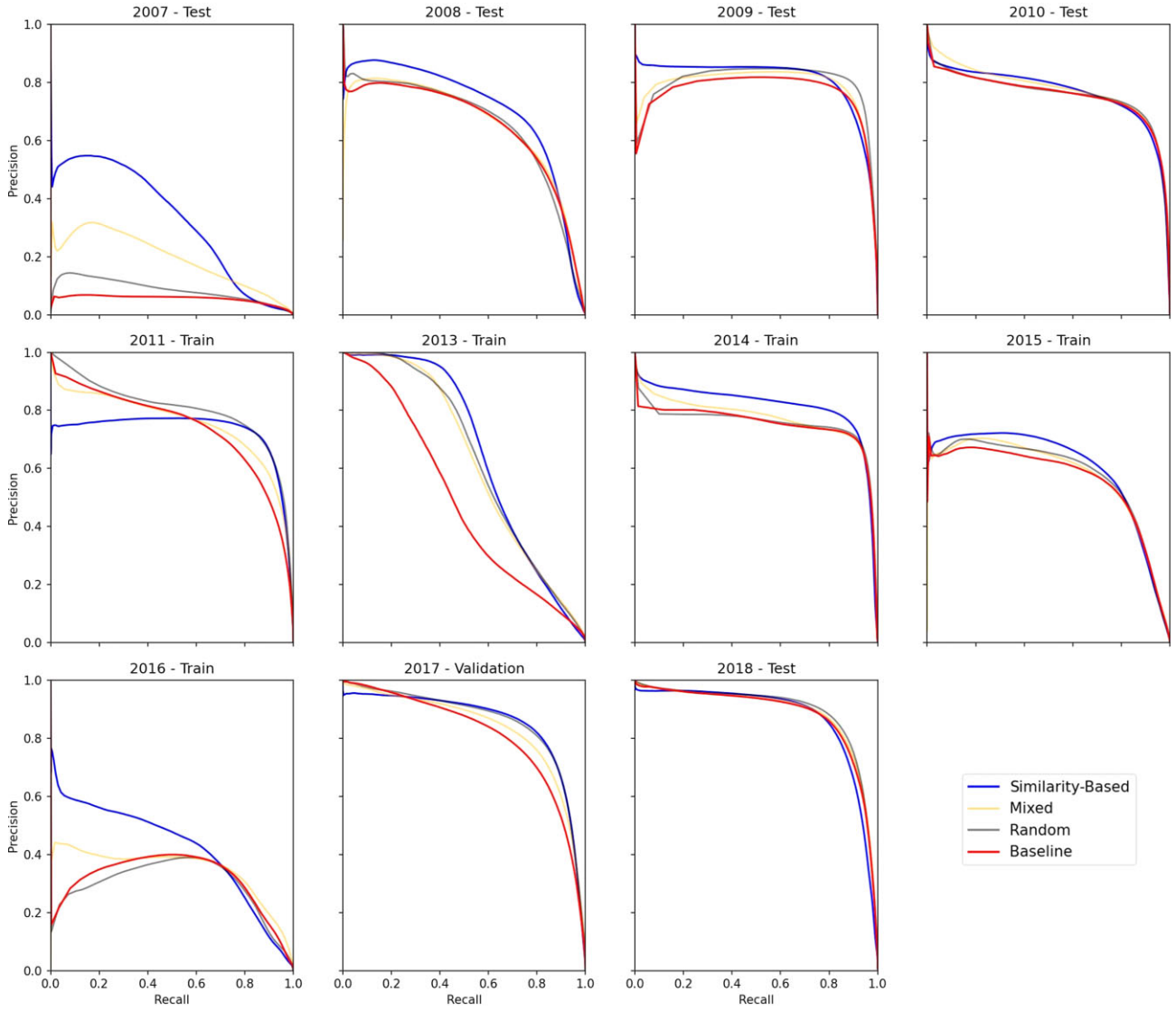


Figure 6. The four models, including random selection of background patches (grey), similarity-based selection (blue), mixed model (yellow), and baseline model (red), are compared across all training, validation, and test years. The performance improvement is particularly evident in the years 2007, 2013, and 2016.

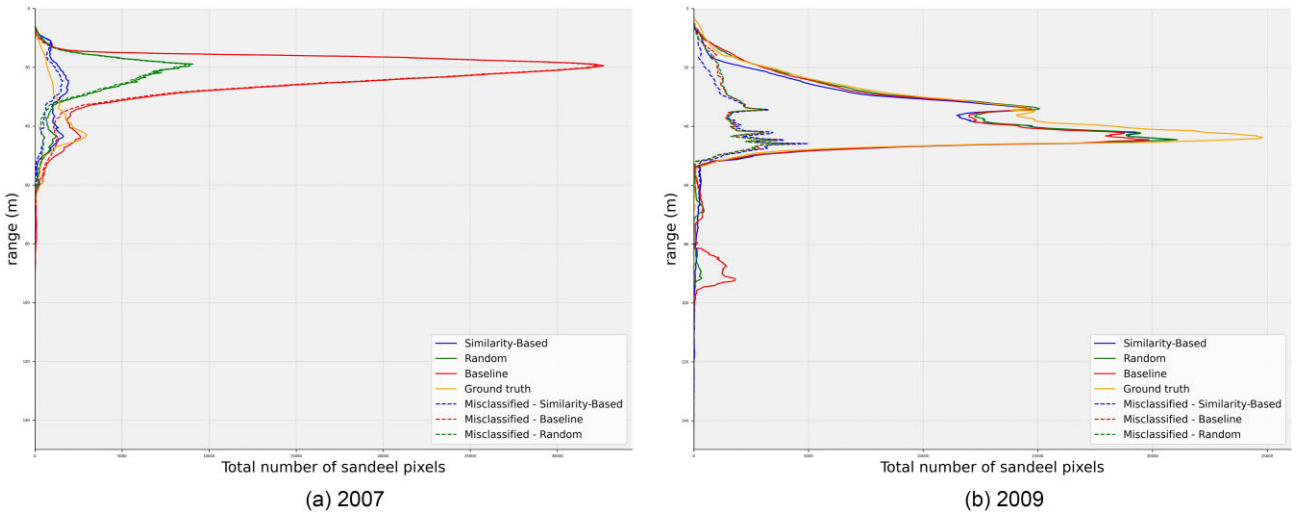


Figure 7. The distributions of "sandeel" pixels by depth for the example test years 2007 (a) and 2009 (b) are visualized. The total number of ground truth pixels count (orange), the count based on the similarity-based model (blue), the count from the random model (green), and the count from the baseline model (red) are visualized. The dashed curves denote the number of misclassified pixels from the similarity-based, random, and baseline models. Notably, the similarity-based model effectively reduces the misclassification of "background" pixels as "sandeel" particularly close to the sea surface.

the test years of 2009, 2010, and 2018, with improvements of 0.0313, 0.0113, and 0.0127, respectively. It is, however, important to note that in those particular years, all the models performed well. On the other side, the similarity-based model yields higher F1 scores compared to the random model in the test years of 2007 and 2008, with improvements of 0.2516 and 0.0373, respectively. Across all the training and validation years, the similarity-based model consistently outperforms the random model, resulting in an F1 score increase ranging from 0.0025 to 0.0377. The random model performed better than the baseline model in all but one year. This suggests that using handcrafted patches is not particularly beneficial. A strategy considering only the pixel classes as patch classes along with applying the proposed sampling methodology, can be effective in enhancing model performance. In addition, since the random undersampling is much less computationally expensive, it is preferable to the baseline sampling.

The random model performed better than the mixed model in all training, validation, and test years except in the test year of 2007. In 2007, the mixed model has a good improvement compared to the random model, with ∼0.1389 higher F1 score. For all other survey years, the random model gives F1 score improvements ranging from 0.0435 to 0.0002, with an average improvement of 0.0131. It was expected that the performance of the mixed model would be between the performance of the random and similarity-based models. However, the random model is mostly performing better than the mixed model. This indicates that when the combination of similarity-based and random sampling is used, the variation between the selected patches for training is not adequately satisfied. For the test years 2009, 2010, and 2018, the mixed model performed slightly better than the similarity-based model but not better than the random model for these years. It is important to note that the mixed model is performing better than the baseline model in all the test, validation, and training years except the training years of 2014 and 2016.

The precision–recall curves (c.f. Figure 6) provide further insights into the performance of the models. The similarity-based model demonstrates higher precision compared to the baseline model in multiple years, including the test years of 2007, 2009, and the training years of 2014 and 2016. This indicates a higher proportion of correctly classified positive instances, i.e. "sandeel" class, among all predicted positive instances. In other words, the samples from the "background" class are classified more precisely. Moreover, when compared to the random and mixed models, the precision–recall curves consistently show that the similarity-based model achieves higher precision values. This shows that the similarity-based model is more accurate in identifying positive instances, resulting in fewer false positives. These findings support the earlier observations based on maximized F1 scores.

Although the similarity-based model consistently outperforms the baseline model, it is important to note that performance varies across different years. In the training year of 2011, for example, while the similarity-based model achieves a higher maximized F1 score (0.7809) compared to the baseline model (0.7114), it exhibits lower precision values at lower threshold levels. This observation demonstrates that the acoustic data from different years have different properties. We also note that the number of not-annotated fish schools may lead to variations in model performance, which seems to be the case for 2011.

We investigated at which echogram range (depth) the similarity-based model improved the "sandeel" pixel predictions compared to the baseline and random models. When considering the counts of ground truth "sandeel" pixels and the numbers of predicted "sandeel" pixels from all these models, we see that there is a high number of misclassifications from the baseline model in the upper water column. Compared to the baseline model, the random model improves the predictions close to the sea surface. However, it still gives a higher number of misclassifications in the upper water column compared to the similarity-based model. This pattern is particularly clear in the years of a poor baseline model performance (c.f. Figure 7a), and less so for the years where the difference between the models is less significant (c.f. Figure 7b). This shows that the similarity-based model reduces misclassifications in the upper field, where plankton layers are known to cause challenges for the baseline model.

A notable difference can be observed in the predictions of the baseline, random, and similarity-based models for a specific region (c.f. Figure 8). The baseline and random models predict the dense layer in the background close to the surface as "sandeel", while the similarity-based model accurately classifies this field as "background". This example prediction highlights the improved performance of the similarity-based model in correctly classifying background structures compared to the baseline and random models. This was one of the main objectives in proposing similarity-based sampling. However, it is also important to note that the similarity-based model, like the baseline and random models, encounters challenges in accurately identifying some parts of the fish schools, as evident from the predictions (c.f. Figure 8e). Although the similarity-based model fails to identify certain portions of the fish schools, the baseline and random models miss an even larger number of fish school parts.

Discussions

This paper addresses the problem of class imbalance in acoustic data through the introduction of an adaptable undersampling approach particularly relevant to semantic segmentation problems. The proposed methodology utilizes state-of-the-art deep learning techniques and ensures a systematic approach to data balancing, avoiding ad hoc methods. A strength of the proposed undersampling approach is that it is data-driven and only uses the pixel classes defined by the annotations. By effectively tackling class imbalance, we contribute to the field of ATC by improving the state-of-the-art deep learning method for semantic segmentation.

Dealing with sampling methodology is crucial when applying deep learning methods to acoustic data, which is characterized by its large size and substantial class imbalance. Although the baseline study (Brautaset *et al.*, 2020) manually balanced the training data by focusing on the "background" class containing seabed samples, it did not account for near-surface and other potential structures. The absence of these in the training dataset led the network to misclassify them. To detect the most informative subsets from the "background" class, our approach identifies the most sandeel-like regions using NearMiss undersampling method, and extracts input patches centred around these regions. This method enhances the network's training and improves model performance, particularly in near-surface areas with dense plankton layers.

The proposed similarity-based sampling noticeably improved the predictions of the network in certain years, such as 2007 from the test set, and 2013 and 2016 from the training set, when the baseline model's performance was poor. For these specific years, the classification of acoustic targets was particularly challenging due to the presence of numerous structures that shared similar backscattering intensity patterns with "sandeel", such as zooplankton or other unidentified structures.

Another finding arising from this research is that the random undersampling method is also useful for deep semantic segmentation models on ATC given that the corresponding model yields better F1 scores in three out of the five test years. It shows an improvement in F1 score compared to the similarity-based model with increments of 0.0313, 0.0113, and 0.0127 for the test years of 2009, 2010, and 2018, respectively. However, it is important to note that the similarity-based model achieves a more substantial improvement in the F1 score for the test years of 2007 and 2008, with increments of 0.2516 and 0.0373, respectively. Moreover, the similarity-based model is better at classifying non-bottom sandeel-like structures, compared to the random model (c.f. Figures 7 and 8). In practice, during manual data scrutiny, the seabed can be separated from the fish since the acoustic properties are usually very different. However, backscatter signals that are similar to the foreground class are typically more challenging. This is where the similarity-based model has its advantage.

The proposed similarity-based sampling approach is general and has the potential to improve the performance of other deep learning models for ATC that face class imbalance. However, when the fish has similar backscatter signals as the seabed, the seabed removal during patch selection should not be performed. Choi *et al.* (2021) randomly undersampled the majority class, in their case the background patches, to achieve class balance. They aim to first cluster the extracted and labelled patches, then classify them using a semi-supervised deep learning framework. For their specific case, the similarity-based sampling method can be applied by selecting representative patches to further improve their deep clustering objective. Vohra *et al.* (2023) used Dice Loss and Focal Loss to address the class imbalance for their deep learning-based semantic segmentation problem, i.e. the automatic detection of underwater discrete scatterers (single marine organisms). In their case, they have 571×1200 pixels of echograms and they did not have to sample from the echograms. Similarity-based sampling could be accomplished by first categorizing echograms according to their background features and prioritizing the echograms with more foreground-like background attributes. Rousseau *et al.* (2022) used undersampling on the majority class (juvenile salmon) for the classification step, and it could be valuable to extend the similarity-based sampling method to classification problems as well. Lastly, in the study by Choi *et al.* (2023), class-rebalancing weights were introduced for the deep semi-supervised semantic segmentation model, and the similarity-based sampling method could complement this approach by ensuring that the selected background samples are sufficiently similar to the foreground class. While this proof-of-concept primarily focuses on ATC and monitoring schools of sandeel, the approach can also be applied to similar problems involving large datasets that require sampling techniques for training purposes, such as satellite image segmentation (Khryashchev *et al.*, 2018) and seismic data segmentation (Birnie *et al.*, 2021).

One limitation of our proposed sampling methodology is that it can inadvertently include sandeel fish schools that are not annotated as "sandeel" within the "background" class. We observed, similarly to Brautaset *et al.* (2020), the presence of fish schools that were either incompletely annotated or not annotated at all. As shown in Figure 4a, these areas are identified by our algorithm, and the corresponding patches are provided to the neural network. Since these regions are considered "background", the network is trained to classify them as "background" rather than "sandeel". As a result, the random model that does not prioritize sandeel-like "background" may achieve slightly better performance in this regard. However, it is important to note that despite this weakness, our proposed sampling strategy substantially improves the performance of the model.

To overcome the difficulties that arise from annotation quality, future studies could focus on handling annotation uncertainty. Applying active learning, in which the uncertainty

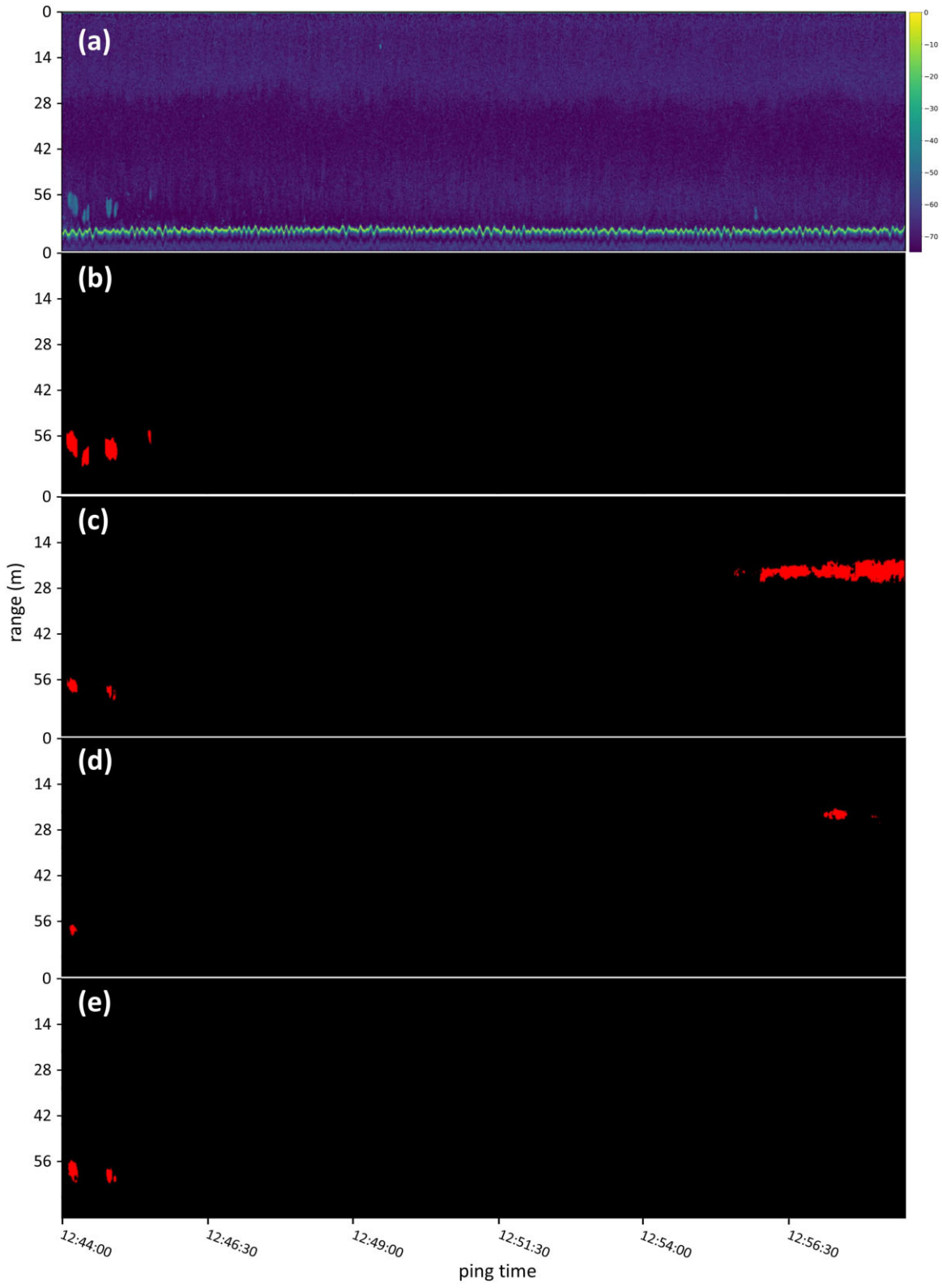


Figure 8. The illustration depicts the comparison of models on the example echogram from 200 kHz (a). It includes the ground truth annotations (b), the predictions of the baseline model (c), the predictions of the random model (d), and the predictions from the similarity-based model (e). The "sandeel" class is visualized in red, while the "background" class is represented in black. The baseline and random models show false positives near the sea surface, where dense plankton layers exist. In contrast, the similarity-based model accurately classifies these parts as "background".

in the labels is discovered and the learning system can engage with a user (or other information sources) to re-label current data points (Yang *et al.*, 2017), is one potential option. By providing more accurate labels, active learning could increase the effectiveness of the deep learning models performed on the acoustic data. Another future direction could be to create an abstract vector representation for the acoustic data using a self-supervised (or unsupervised) technique. The key benefit is that classifiers can be trained fast (even interactively) on top of the embedding, freeing us from relying on annotations, which are technically challenging to obtain and likely to be erroneous.

We have proposed an adaptable approach to select data from echograms for training deep learning models. The approach is based on a similarity-based undersampling methodology, which identifies the most foreground-like regions. It offers a solution to effectively address the challenge of handling class imbalance in acoustic data and contributes to the development of ATC methods.

Author contributions

AP: Conceptualization and planning, Methodology, Data processing, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Resources, Visualization. AO: Conceptualization and planning, Methodology, Investigation, Writing – original draft, Writing – review & editing, Resources, Supervision. IU: Data processing, Software, Writing – review & editing, Resources. NOH: Conceptualization and planning, Methodology, Investigation, Writing – original draft, Writing – review & editing, Resources, Supervision, Funding acquisition, Project administration.

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 Marine Acoustic Abundance Estimation and Backscatter Classification (CRIMAC) project (no. 309512). Anna Oleynik is funded through the Academia agreement between Equinor and the University of Bergen.

Data availability

Data available on request: the data underlying this article will be shared on reasonable request to the corresponding author.

References

Abdi, H.and and Williams, L. J. 2010. Principal component analysis. Wiley Interdisciplinary Reviews Computational Statistics, 2: 433–459.

Bao, L., Juan, C., Li, J., and Zhang, Y. 2016. Boosted near-miss undersampling on SVM ensembles for concept detection in large-scale imbalanced datasets. Neurocomputing, 172: 198–206.

Barandela, R., Valdovinos, R. M., Sánchez, J. S., and Ferri, F. J. 2004. The imbalanced training sample problem: under or over sampling? *In* Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, SSPR 2004 and SPR 2004, Lisbon, Portugal, August 18–20, 2004 Proceedings, pp. 806–814. Springer, New York, NY.

Beyan, C.and and Fisher, R. 2015. Classifying imbalanced data sets using similarity based hierarchical decomposition. Pattern Recognition, 48: 1653–1672.

Birnie, C., Jarraya, H.and and Hansteen, F. 2021. An introduction to distributed training of deep neural networks for segmentation tasks with large seismic data sets. Geophysics, 86: KS151–KS160.

Blackwell, R. E., Harvey, R., Queste, B. Y., and Fielding, S. 2020. Colour maps for fisheries acoustic echograms. ICES Journal of Marine Science, 77: 826–834.

Błaszczyński, J.and and Stefanowski, J. 2015. Neighbourhood sampling in bagging for imbalanced data. Neurocomputing, 150: 529–542.

Brautaset, O., Waldeland, A. U., Johnsen, E., Malde, K., Eikvil, L., Salberg, A.-B., and Handegard, N. O. 2020. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. ICES Journal of Marine Science, 77: 1391–1400.

Buda, M., Maki, A., and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106: 249–259.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321–357.

Choi, C., Kampffmeyer, M., Handegard, N. O., Salberg, A.-B., and Jenssen, R. 2023. Deep semisupervised semantic segmentation in multifrequency echosounder data. IEEE Journal of Oceanic Engineering, 48: 384–400.

Choi, C., Kampffmeyer, M., Handegard, N. O., Salberg, A.-B., Brautaset, O., Eikvil, L., and Jenssen, R. 2021. Semi-supervised target classification in multi-frequency echosounder data. ICES Journal of Marine Science, 78: 2615–2627.

Foote, K. G. 1983. Linearity of fisheries acoustics, with addition theorems. The Journal of the Acoustical Society of America, 73: 1932–1940.

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.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42: 463–484.

Gastauer, S., Fässler, S. 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.

Haralabous, J.and and Georgakarakos, S. 1996. Artificial neural networks as a tool for species identification of fish schools. ICES Journal of Marine Science, 53: 173–180.

Hasib, K. M., Iqbal, M., Shah, F. M., Mahmud, J. A., Popel, M. H., Showrov, M., Hossain, I. *et al.* 2020. A survey of methods for managing the classification and solution of data imbalance problem. Journal of Computer Science, 16: 1546–1557.

Hu, S., Liang, Y., Ma, L., and He, Y. 2009. Msmote: Improving classification performance when training data is imbalanced. *In* 2009 Second International Workshop on Computer Science and Engineering, 2. IEEE, New York, NY. pp. 13–17.

Japkowicz, N. 2000. The class imbalance problem: significance and strategies. Proceedings of the International Conference on Artificial Intelligence, 56: 111–117.

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., Rieucau, 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 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.

Khryashchev, V., Ivanovsky, L., Pavlov, V., Ostrovskaya, A., and Rubtsov, A. 2018. Comparison of different convolutional neural network architectures for satellite image segmentation. *In* 2018 23rd Conference of Open Innovations Association (FRUCT), pp. 172–179. IEEE, New York, NY.

Kloser, R., Ryan, T., Sakov, P., Williams, A., and Koslow, J. 2002. Species identification in deep water using multiple acoustic frequencies. Canadian Journal of Fisheries and Aquatic Sciences, 59: 1065–1077.

Korneliussen, R. J. 2000. Measurement and removal of echo integration noise. ICES Journal of Marine Science, 57: 1204–1217.

Korneliussen, R. J. 2002. Analysis and presentation of multi-frequency echograms. Ph.D. thesis, University of Bergen, Bergen.

Korneliussen, R. J. 2018. Acoustic target classification. International Council for the Exploration of the Sea (ICES), Copenhagen.

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.

Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5: 221–232.

Kubat, M. and Matwin, S. 1997. Addressing the curse of imbalanced training sets: one-sided selection. The International Conference on Machine Learning (ICML),97 179–186

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

Lin, W.-C., Tsai, C.-F., Hu, Y.-H., and Jhang, J.-S. 2017. Clustering-based undersampling in class-imbalanced data. Information Sciences, 409: 17–26.

Lurton, X. 2002. An Introduction to Underwater Acoustics: Principles and Applications, 2. Springer, New York, NY.

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.

Mani, I.and and Zhang, I. 2003. kNN approach to unbalanced data distributions: a case study involving information extraction. Proceedings of Workshop on Learning from Imbalanced Datasets, ICML, 126: 1–7.

Marques, T. P., Cote, M., Rezvanifar, A., Albu, A. B., Ersahin, K., Mudge, T., and Gauthier, S. 2021a. 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. 4378–4387. IEEE, New York, NY.

Marques, T. P., Rezvanifar, A., Cote, M., Albu, A. B., Ersahin, K., Mudge, T., and Gauthier, S. 2021b. Detecting marine species in echograms via traditional, hybrid, and deep learning frameworks. *In* 2020 25th International Conference on Pattern Recognition (ICPR), pp. 5928–5935. IEEE, New York, NY.

Mumuni, A.and and Mumuni, F. 2022. Data augmentation: a comprehensive survey of modern approaches. Array, 16: 100258.

Ng, W. W., Hu, J., Yeung, D. S., Yin, S., and Roli, F. 2014. Diversified sensitivity-based undersampling for imbalance classification problems. IEEE Transactions on Cybernetics, 45: 2402–2412.

Ordonez, A., Utseth, I., Brautaset, O., Korneliussen, R., and Handegard, N. O. 2022. Evaluation of echosounder data preparation strategies for modern machine learning models. Fisheries Research, 254: 106411.

Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y.-G., Ding, K. *et al.* 2019. Trainable undersampling for class-imbalance learning. Proceedings of the AAAI Conference on Artificial Intelligence, 33: 4707–4714.

Redmon, J.and and Farhadi, A. 2017. Yolo9000: better, faster, stronger. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271. IEEE, New York, NY.

Reid, D. 2000. Cooperative research report on echo trace classification. ICES, Copenhagen.

Ren, S., He, K., Girshick, R., and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28.

Rezvanifar, A., Marques, T. P., Cote, M., Albu, A. B., Slonimer, A., Tolhurst, T., Ersahin, K. *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 (last accessed 7 May 2023).

Ronneberger, O., Fischer, P., and Brox, T. 2015. U-net: convolutional networks for biomedical image segmentation. *In* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015 Proceedings, Part III 18, pp. 234–241. Springer, New York, NY.

Rousseau, S., Gauthier, S., Neville, C., Johnson, S., and Trudel, M. 2022. Acoustic classification of juvenile pacific salmon (*Oncorhynchus* spp) and pacific herring (*Clupea pallasii*) schools using random forests. Frontiers in Marine Science, 9: 857645.

Salau, A. O., Markus, E. D., Assegie, T. A., Omeje, C. O., and Eneh, J. N. 2023. Influence of class imbalance and resampling on classification accuracy of chronic kidney disease detection. Mathematical Modelling of Engineering Problems, 10: 48–54.

Simmonds, J.and and MacLennan, D. N. 2008. Fisheries acoustics: theory and practice. John Wiley and Sons, New York, NY.

Sowah, R. A., Agebure, M. A., Mills, G. A., Koumadi, K. M., and Fiawoo, S. Y. 2016. New cluster undersampling technique for class imbalance learning. International Journal of Machine Learning and Computing, 6: 205–214.

Summers, C.and and Dinneen, M. J. 2019. Improved mixed-example data augmentation. *In* 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1262–1270. IEEE, New York, NY.

Talaat, F. Mand and ZainEldin, H. 2023. An improved fire detection approach based on YOLO-v8 for smart cities. Neural Computing and Applications, pp.1–16. Springer.

Tomek, I. 1976. Two modifications of CNN. IEEE Transactions on Systems Man and Communications, 6: 769–772.

Ullah, A., Javaid, N., Asif, M., Javed, M. U., and Yahaya, A. S. 2022. Alexnet, adaboost and artificial bee colony based hybrid model for electricity theft detection in smart grids. IEEE Access, 10: 18681–18694.

Van Hulse, J., Khoshgoftaar, T. M., and Napolitano, A. 2007. Experimental perspectives on learning from imbalanced data. *In* Proceedings of the 24th International Conference on Machine learning, pp. 935–942. ICML, Oregon, OR.

Vohra, R., Senjaliya, F., Cote, M., Dash, A., Albu, A. B., Chawarski, J., Pearce, S. *et al.* 2023. Detecting underwater discrete scatterers in echograms with deep learning-based semantic segmentation. *In* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 375–384. IEEE, New York, NY.

Wang, Z., She, Q., and Ward, T. E. 2021. Generative adversarial networks in computer vision: a survey and taxonomy. ACM Computing Surveys (CSUR), 54: 1–38.

Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. 2016. Understanding data augmentation for classification: when to warp? *In* 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–6. IEEE, New York, NY.

Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D. Z. 2017. Suggestive annotation: a deep active learning framework for biomedical image segmentation. *In* Medical Image Computing and Computer Assisted Intervention—MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017 Proceedings, Part III 20, pp. 399–407. Springer, New York, NY.

Zhou, Z.-H.and and Liu, X.-Y. 2010. On multi-class cost-sensitive learning. Computational Intelligence, 26: 232–257.

Handling editor: Ahmad Salman

Received: 13 July 2023; **Revised:** 28 September 2023; **Accepted:** 3 October 2023