LIMNOLOGY and OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 2024

doi: 10.1002/lom3.10611

Developing a hybrid model with multiview learning for acoustic classification of Atlantic herring schools

Yawen Zhang,[1]* **Carrie C. Wall**,[2] **J. Michael Jech**,[3] **Qin Lv**[1]

[1]Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, USA
[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder; NOAA National Centers for Environmental Information, Boulder, Colorado, USA
[3]NOAA Northeast Fisheries Science Center, Woods Hole, Massachusetts, USA

Abstract

Advances in active acoustic technology have outpaced the ability to process and analyze the data in a timely manner. Currently, scientists rely on manual scrutiny or limited automation to translate acoustic backscatter to biologically meaningful metrics useful for fisheries and ecosystem management. The National Oceanic and Atmospheric Administration Northeast Fisheries Science Center has monitored the Atlantic herring population in the Gulf of Maine and Georges Bank since 1999 due to the stocks' important economic and ecological role for the commercial lobster industry. Manual scrutinization to identify Atlantic herring schools from the water column sonar data is time-consuming and impractical for large-scale studies. To automate this process, a hybrid model with multiview learning was proposed for automatic Atlantic herring school detection, which consists of two steps: (1) region-of-interest (ROI) detection and (2) ROI classification. The ROI detection step was designed to detect school-like objects, and the ROI classification step was designed to distinguish Atlantic herring schools from other objects. The co-training algorithm was employed for multiview learning as well as semi-supervised learning. Within this framework, single-view vs. multiview learning and supervised vs. semi-supervised learning were evaluated and compared. Our results showed that multiview learning can improve the performance of the hybrid model in Atlantic herring school detection, and the utilization of unlabeled data is also helpful when the training set is small. The best-performed model achieved an *F*1-score of 0.804. This new framework provides an efficient and effective tool for automatic Atlantic herring school detection.

To date, marine scientists rely on manual scrutiny or limited automation to analyze acoustic data by delineating acoustic signals of species of interest. Such manual or semi-automatic methods are time-consuming, impractical for large-scale studies, and very difficult to reproduce without domain expertise. In addition, manual annotations may include missing or incomplete data, whose quality is difficult to assess (Brautaset et al. 2020). To address the challenge of efficiently analyzing large, complex acoustic survey data (Wall et al. 2016), as well as provide a complementary approach to existing annotation processes, automatic approaches for acoustic target detection are becoming increasingly important (Beyan and Browman 2020).

The objective of automatic acoustic target detection can be summarized in two steps: localization and classification. The first step localizes the region of interest (ROI) and the second step differentiates the target from other objects. Previous studies have used Echoview (Myriax Pty, Ltd) to detect fish schools and then classify based on a variety of school-related metrics (Jech and Sullivan 2014; Proud et al. 2020). An ideal, fully automated system for acoustic target detection would encompass all the necessary steps in a single workflow (Malde et al. 2020). Recent studies have explored machine learning (Fallon et al. 2016; Korneliussen et al. 2016; Proud et al. 2020) and deep learning (Rezvanifar et al. 2019; Brautaset et al. 2020; Porto Marques et al. 2021) techniques to solve this problem. In general, there are two categories of models: hybrid and end-to-end. Hybrid models conduct object localization and classification separately resulting in two separate modules. For example, Rezvanifar et al. developed a two-step framework consisting of a ROI extractor and a ROI classifier to detect herring schools from acoustic data (Rezvanifar et al. 2019). End-to-end models conduct object localization and classification in a unified model. For example, Marques et al. applied the state-of-the-art object detection model You Only Look Once

*Correspondence: yawen.zhang@colorado.edu

(YOLO) to detect herring schools (Porto Marques et al. 2021), where YOLO (Redmon et al. 2016) is a representative example of end-to-end models. In hybrid models, the only trainable part is often the ROI classifier, which is "light-weight" compared with end-to-end models, and thus requires fewer samples to train. Marques et al. conducted an extensive comparison of hybrid and end-to-end models and showed that hybrid models can achieve comparable performance (with a better *F*1-score) to end-to-end models (e.g., YOLO, Faster R-CNN; Girshick 2015; Porto Marques et al. 2021).

In addition to model architecture, contextual features (i.e., covariates), such as depth, geographic location, and environmental conditions, can add useful information and improve the effectiveness of the models. A previous study demonstrated that adding context metadata, such as sample depth and location, improved plankton image classification performance (Ellen et al. 2019). To the best of our knowledge, such contextual information has not been well utilized in object detection models, especially for acoustic target detection. For end-to-end object detection models, the inputs are fixed size images and acoustic echograms lose their contextual information once transformed into images. Hybrid models are more flexible than end-to-end models in terms of their structure because they have separate modules, which provide the opportunity to incorporate such information.

Therefore, this study aims to develop a hybrid model with multiview learning. Multiview learning uses multiple views to improve model performance (Zhao et al. 2017). The method adopted here is called co-training (Blum and Mitchell 1998), a well-known multiview learning method that has been applied in many different applications (Zhu et al. 2014; Li et al. 2023). In a co-training algorithm, two classifiers are trained from two different views that predict the data for each other, and each view consists of a set of features. As such, the training set is enlarged with high-confidence predictions from both classifiers. The major advantages of using the co-training algorithm for acoustic target classification include: (1) contextual information such as depth in the water and geographic location can be included as another view that is complementary to the visual view, which are mainly visual features of acoustic targets, and (2) a semi-supervised machine learning approach (Hastie et al. 2009) that uses unlabeled data in the training process, which is beneficial when only a small amount of labeled data are available.

We use acoustic survey data as test of these methods. Acoustic surveys are widely conducted to study underwater species and assess their abundance. In 1999, the National Oceanic and Atmospheric Administration (NOAA) Northeast Fisheries Science Center (NEFSC) began acoustic-trawl surveys to determine the annual biomass estimates of the Atlantic herring (*Clupea harengus*) population in the Gulf of Maine and Georges Bank because of the stock's economic and ecological importance as bait for the commercial lobster industry (Brandt and McEvoy 2006; NOAA Northeast Fisheries Science Center 2012, 2018; Jech and Sullivan 2014). Atlantic herring is widely distributed in the Northwest Atlantic Ocean and is a key prey species for many predators (Jørstad et al. 1991). They are fall spawners who migrate from their feeding grounds to Georges Bank to spawn every fall season (Reid et al. 1999; Stephenson et al. 2009). To identify Atlantic herring schools from the NEFSC surveys, Jech et al. manually outlined regions with Atlantic herring and used Echoview to detect Atlantic herring schools (Jech and Sullivan 2014). The surveys employed a multiple frequency echosounder to collect acoustic data from both biological and nonbiological objects in the water column. The echosounder's multiple frequencies result in a spectral profile that helps scientists identify trophic levels, and species of interest when combined with the associated trawl data (Horne 2000; Korneliussen 2018). However, extracting individual species targets from backscatter remains challenging because species-level information is not always available (Trenkel and Berger 2013), and some species are difficult to distinguish from others with similar morphology and behavior (Korneliussen et al. 2016).

The primary objective of this study was to develop a hybrid model for acoustic classification of Atlantic herring schools. The model's framework was designed to simplify the NEFSC fisheries acoustic expert's original process to annotate these schools, reduce manual effort, and provide comparable results. The model consisted of two key steps: ROI detection and ROI classification. The secondary objective of this study was to explore the multiview machine learning method for species classification, in particular, the co-training algorithm. The co-training algorithm was implemented with different feature extraction approaches. The effectiveness of this algorithm to identify Atlantic herring schools was demonstrated through comparison with expert annotations.

Materials and background

Data sources

Multifrequency echosounder data collected by the NOAA NEFSC were used in this study. The survey series from 2009 to 2022 were collected by the NOAA Ship *Henry B. Bigelow* during both daytime and nighttime as part of the NEFSC's bottom trawl survey (Politis et al. 2014). The survey design is stratified random where bottom trawl locations are randomly selected within strata, which are based on bathymetry. All acoustic data collected in-transit between trawl deployments and during deployments were used in this analysis. Acoustic data consisted of multifrequency Simrad EK60 echosounder narrowband (i.e., continuous wave, CW) data at 18, 38, 70, 120, and 200 kHz. The echosounders were calibrated following standard procedures (Foote 1987) before each survey. Pulse duration was set to 1 ms for all frequencies during operation. See Jech et al. (2000), Jech and Stroman (2012), and Jech and Sullivan (2014) for more details about the NEFSC acoustic survey data. All the calibrated, multiple frequency single-

beam data collected from 1999 to 2022 are archived at the NOAA National Centers for Environmental Information (NCEI) (2020) and publicly accessible through Amazon Web Services (AWS).[1]

We selected acoustic survey data collected from September 25 to November 14 (39 days, 1192 echograms) in 2019 as a representative test set (NOAA Northeast Fisheries Science Center 2019). Because the goal of this study was to investigate and evaluate methods, we opted not to apply the data to the full time series so that we could focus on finding a successful model. The vessel trajectories are shown in Figure 1. Data prior to the first encounter of Atlantic herring during each survey (i.e., the southernmost extent of the survey) were scrutinized but not included as there were no Atlantic herring schools present. The survey data were processed and annotated by the NEFSC fisheries acoustics expert (861 echograms with annotations). Echosounder raw recordings were post-processed as multifrequency volume backscatter (Sv data) using an open-source Python library called PyEcholab.[2] The echograms are 2454 (time dimension) × 2613 (depth dimension) in pixels. All echograms were preprocessed by aligning pings in the time/distance domain across the four frequency components. The seabed detected by the EK60 echosounder software and documented in .bot files were used to remove data under the seafloor. To minimize the effect of surface bubbles and erroneous seafloor detection, data shallower than 10 m or data deeper than 1 m above the seafloor were excluded. To smooth and diminish stochastic local variation, a 3 × 3 median filter was applied to each frequency component.

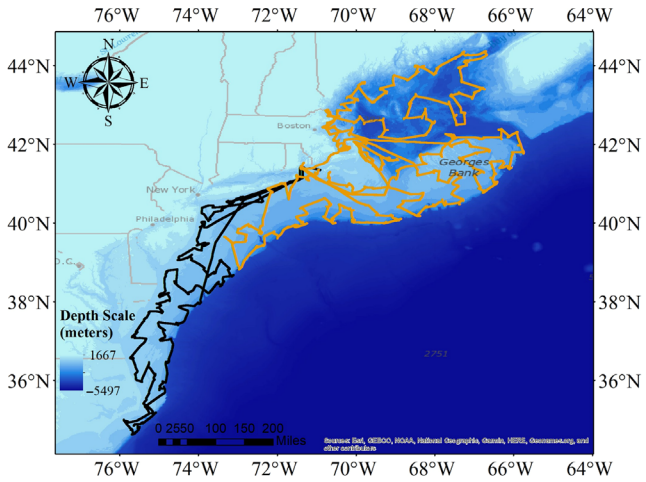

Fig. 1. The NOAA Ship *Henry B. Bigelow* acoustic-trawl survey track in 2019 overlaid on the ETOPO1 1 Arc-Minute Global Relief Model (NOAA National Geophysical Data Center 2009). Orange denotes the survey data processed and annotated by the NEFSC fisheries acoustics expert.

Human annotations

Annotation process

The workflow used by the NEFSC fisheries acoustics expert to extract Atlantic herring schools (in this paper, we define "school" as any aggregation of Atlantic herring including schools and shoals) from the NEFSC acoustic survey data consisted of five steps: (i) *Bottom detection and error correction*: in this step, an automated seabed detection algorithm was applied using Echoview. The results were then manually quality controlled to remove erroneously labeled seabed echos. This is an important step as the inclusion of seabed echoes is the largest source of error in abundance estimates. (ii) *Manual nonbiological object outlining*: while scrutinizing the seabed detection, scattering features that were not of biological origin, for example, echoes from conductivity–temperature–depth (CTD) recorders deployed during survey, surface bubbles, and noise artifacts, were identified and labeled (Fig. 2). (iii) *Acoustic classification*: based on the multifrequency single-beam imaging method (Jech and Michaels 2006; Wall et al. 2016), only the scattering that was indicative of gas-bearing targets was retained since herring was a subset of those targets. (iv) *Manual target outlining and masking*: This step located the target regions by manually drawing polygon regions circumscribing backscatter that were believed to be due to Atlantic herring. The outlined regions were rather crude and could incorporate scatter from sources outside of Atlantic herring. (v) *Fish school detection*: as the final step, the outlined regions were used to mask the original data, and only regions with Atlantic herring were retained. Echoview's "Schools Detection Module" was applied to the retained regions using the 38-kHz data to more precisely detect Atlantic herring schools. The 38-kHz data were selected because this frequency is used to generate abundance estimates for Atlantic herring assessments. School detection was performed using Shoal Analysis and Patch Estimation System (SHAPES) (Barange 1994; Coetzee 2000), which requires a number of parameters to be set. The goal of this step was to group all the contiguous bins and remove all the speckles. The minimum total school height and length were all set to 4 m, the minimum candidate length was set to 1 m, and the minimum candidate height was set to 2 m. The minimum vertical linking distance was set to 2 m, and the maximum horizontal linking distance was set to 20 m (Jech and Stroman 2012). More details about Echoview's school detection module can be found here.[3] Essentially, each step further eliminates backscatter that is not associated with Atlantic herring schools based on empirical knowledge (Jech and Sullivan 2014).

Based on this annotation process, four classes of annotations were created: positive target "Atlantic herring school", and negative nontarget "*Clupeid* school," "Krill," and "Noise region" (Table 1). Figure 2 provides examples of annotations.

[1] https://registry.opendata.aws/ncei-wcsd-archive/

[2] https://github.com/CI-CMG/pyEcholab

[3] https://support.echoview.com/WebHelp/Reference/Algorithms/Schools_Detection_Module/Notes_about_schools_detection.htm

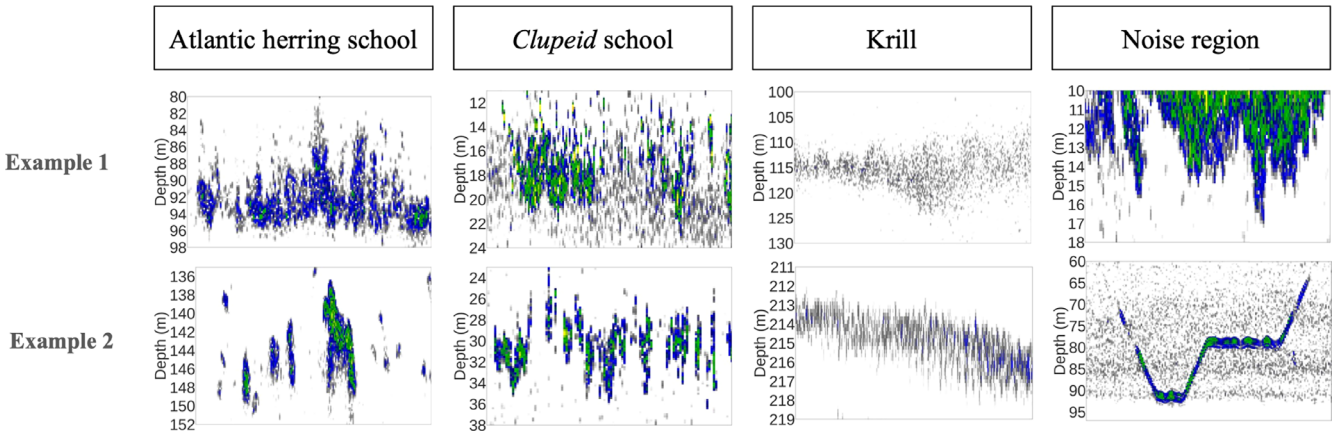

Fig. 2. Examples of annotations, which depict 38 kHz Sv values with a color scale of −66 to 0 dB re 1 m^{-1} (from gray to blue to green). Sv values less than −66 dB re 1 m^{-1} have been excluded and shown as the white background. Noise regions show the echoes from bubbles at the sea surface (top panel) and a CTD rosette (lower panel).

All the school annotations were generated by Echoview's school detection algorithm. "Atlantic herring school" denotes a school with scattering features of Atlantic herring, which is the target class. "Clupeid school" were generated within regions that were outlined as possible herring, which denotes scattering features associated with Clupeid-type fish, such as alewife (*Alosa pseudoharengus*), blueback herring (*Alosa aestivalis*), shad (*Alosa sapidissima*), or menhaden (*Brevoortia tyrannus*). "Krill" denotes scattering features associated with krill (*Euphausia* sp.) that were identified using methods described by Jech et al. (2018). "Noise region" denotes regions of noise, such as CTD rosette echoes and surface bubbles from inclement weather (Shabangu et al. 2014). The annotations were used to assign labels to the ROIs as described in "ROI Labels" section and evaluate the model performance as described in "Evaluation" section.

Annotation characteristics

As shown in Figure 2, Sv classified as "Atlantic herring school" had similar acoustic intensity as "Clupeid school" and "Noise region" at 38 kHz, whereas "Krill" had a much weaker acoustic intensity than the other classes. Some annotations of Clupeid schools were acoustically similar to Atlantic herring but exhibited different depth preferences. Table 1 presents the total number and basic characteristics (e.g., depth, size) of each class. Clupeid schools were mostly abundant in shallow water (depth around 29 m) while Atlantic herring schools were most abundant in deeper water (depth around 117 m). Noise regions had the shallowest median depth of 14 m. Besides differences in depth distribution, Clupeid schools and krill tended to form smaller aggregations than Atlantic herring schools since their median sizes were smaller. It is also worth noting that Clupeid school, krill, and noise regions only form a subset of all the negative (i.e., nontarget) classes. It is important to have an effective classifier that can differentiate Atlantic herring schools from all possible negative classes.

Among Atlantic herring school annotations, there were also variations in terms of their acoustic, geometric, and geographic characteristics. Figure 3 shows the multifrequency responses and aspect ratio (ratio of fish school's length to thickness) of Atlantic herring schools at different depth ranges. In addition, their frequency responses changed with depth. In shallower water, they tended to have stronger responses at lower frequencies (18 and 38 kHz) than higher frequencies (120 and 200 kHz); while in deeper water, this trend gradually reversed. For example, at depth range of 200–250 m, Atlantic herring schools had stronger responses at higher frequencies (120 and 200 kHz) than lower frequencies (18 and 38 kHz). At the same time, as depth increased, the aspect ratio of Atlantic herring school increased, which indicates that schools near the seafloor extended for longer distances than those near the surface. The high variability of aspect ratio (from 0 to 60) also posed some challenge to the detection of Atlantic herring schools since the targets' shape

Table 1. Summary of annotations generated from the NEFSC for 2019 herring survey.

Category	Class name	#Annotations	Median depth (m)	Median size (m^2)
Positive (target)	Atlantic herring school	2140	117	318
Negative (nontarget)	Clupeid school	694	29	172
	Krill	89,470	70	136
	Noise region	159	14	665

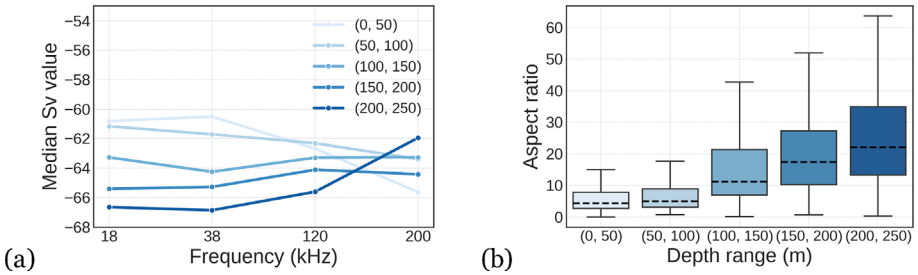

Fig. 3. Depth-dependent characteristics of Atlantic herring schools: (**a**) multifrequency Sv values and (**b**) aspect ratio (the ratio of school's length to thickness) at different depth ranges.

can change significantly. The differences in frequency responses and aspect ratios together demonstrate the depth-dependent characteristics of Atlantic herring schools.

Framework and evaluation

In this study, the hybrid model for Atlantic herring school detection consisted of two key steps: (i) ROI detection and (ii) ROI classification. The first step detects ROIs that are likely to be Atlantic herring schools while the second step distinguishes Atlantic herring schools from other classes. Figure 4 provides a graphical comparison of the original annotation process and the hybrid model with the outputs of each step.

Our new framework implemented entirely in Python substantially reduces the number of steps and manual effort.

ROI detection

Contour detection was employed to detect ROIs (Suzuki and be 1985). Contours are curves that join continuous points along the boundary with the similar color or intensity. Contour detection serves as the basis of object detection (Arbelaez et al. 2010) and is well suited to detect fish school-like objects. The *findContours* function from Python's OpenCV library was used for ROI detection. By applying the same minimum length and thickness constraints as described in "Human Annotations" section, very tiny ROIs were excluded.

This method is nonparametric. The only parameter that impacts ROI detection is the minimum Sv value, which is set to exclude backscatter that is not related to the target. The threshold of −66 dB re 1 m^{-1} has been previously applied to detect Atlantic herring (Jech and Sullivan 2014). In many previous studies (Brautaset et al. 2020; Choi et al. 2021), this threshold was empirically set given the target species. In this study, instead of setting it empirically, varying values of the Sv threshold were applied for ROI detection, ranging from −80 to −54 dB re 1 m^{-1} with an interval of 2 dB re 1 m^{-1}. The optimal Sv threshold for these data was chosen to optimize

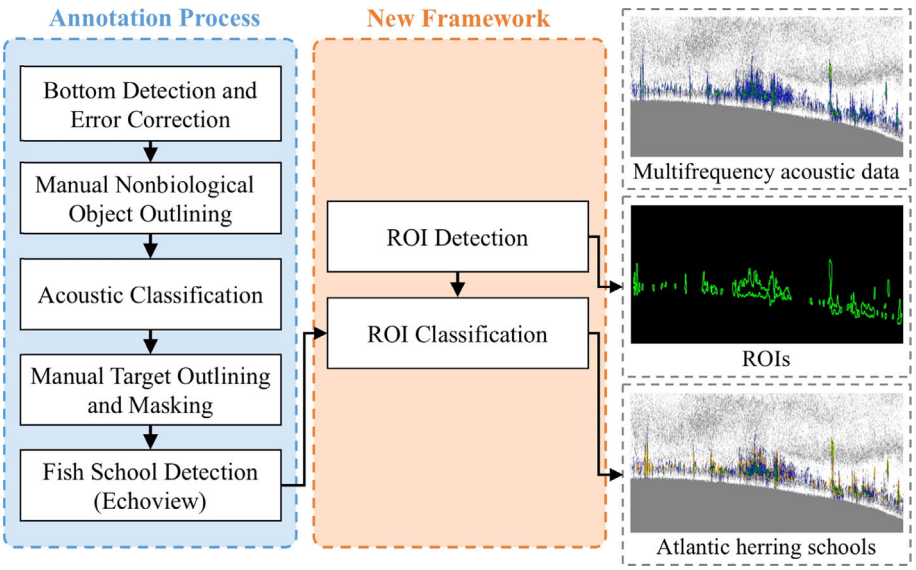

Fig. 4. Comparison of the original, human-in-the-loop annotation process (left) and the hybrid model (center) for Atlantic herring school detection. Examples of unlabeled acoustic data (top right), ROI detection result (middle right), and ROI classification results (bottom right) are shown in the right panel.

the model performance (*see* "ROI Detection: Sv Threshold" section for detailed results).

ROI classification

The detected ROIs include Atlantic herring schools as well as other objects like Clupeid schools, noise, and seabed echoes. In this step, binary classifiers were trained to distinguish Atlantic herring schools from other objects. The co-training algorithm was employed to enable multiview learning for ROI classification and incorporate contextual information.

ROI labels

ROI labels were acquired by overlapping the detected ROIs with human annotations. ROIs that had a significant overlap with annotations got their labels from annotations. Their overlap was measured by the intersection-over-union (IoU) metric (ranging from 0 to 1) that is widely used to measure the performance of object detection methods (*see* the graphical demonstration in Figure 7, in which the "target" is the union of two objects). A high IoU value indicates a good overlap between an ROI and an annotation. The IoU threshold was set to 0.5, indicating that the ROI and annotation must overlap by at least 50% in order to get labeled. The ROIs that did not satisfy this condition would remain unlabeled. Given the available annotations, there were four possible labels for the detected ROIs, including "Atlantic herring school," "Clupeid school," "Krill," and "Noise region." The sum of these four regions did not account for the total Sv of the full echogram since (1) Sv that was below the threshold were not included and (2) there were other nontarget objects beyond the annotated ones.

Training, validation, and test

The acoustic survey data were divided into training, validation, and test sets by a random 60–20–20 split of the 1192 echograms (as shown in Table 2). The training set was used to detect ROIs and train ROI classification models at different Sv thresholds. The validation set was used to evaluate the models' performance with different hyperparameter settings. In this framework, the Sv threshold was also included as a hyperparameter, and its optimal value was selected by the validation set. The test set was used to report the final performance of the whole framework after completing the training process. By overlapping with annotations, ROIs were divided into three groups: *labeled positive*, *labeled negative*, and *unlabeled negative*. *Labeled positive ROIs* were labeled as Atlantic herring school and corresponded to annotated Atlantic herring schools. *Labeled negative ROIs* were labeled as the nontarget classes Clupeid school, Krill, or Noise region and corresponded to annotations of those types. *Unlabeled negative ROIs* were those without any labels and came from echograms without annotations of any type.

The reason to include *unlabeled negative ROIs* is that when human experts labeled samples, they focused more on the positive samples (Atlantic herring school) and less on the negative samples, so they may have only labeled a few negative classes. However, more negative classes beyond the classes labeled may exist in the data, therefore, *labeled negative* and *unlabeled negative ROIs* were combined into *all negative ROIs* to provide a comprehensive picture of the negative samples. Table 2 reports the number of ROIs in each category when applying the Sv threshold of -66 dB re 1 m^{-1}. It is important to mention that the dataset exhibits class imbalance, with a ratio of positive to negative examples being 1 : 22 in the training set. The optimal ratio of negative to positive samples was chosen by optimizing the overall performance on the validation set (*see* "ROI Classification: Ratio of Negative to Positive Samples" section for detailed results).

Multiview learning algorithm

Co-training was employed for multiview learning, which incorporates two complementary views for ROI classification. As shown in Figure 5, different categories of features created two views: View 1 with school-related features, and View 2 with contextual features. Each view can be individually trained to discriminate Atlantic herring schools from other categories. View 1 was trained with features directly related to fish or krill schools, such as their acoustic and geometric characteristics. There are typically two approaches to generate school-related features: one is via feature engineering (also called handcrafted [HC] features), and the other is via the convolutional neural networks (CNN) model. View 2 was trained with contextual features such as depth and location that quantify the geographic location of the fish or krill schools. One assumption of co-training is that these views are conditionally independent. However, in practice that criterion is usually not met

Table 2. Summary of the training, validation, and test datasets used in the models. The number of echograms (#Echograms), number of echograms with annotations (#Annotated echograms), number of ROIs labeled with Atlantic herring schools (#Labeled positive ROIs), number of ROIs labeled with nontarget classes such as Clupeid school, Krill, or Noise region (#Labeled negative ROIs), and the number of labeled negative and unlabeled negative ROIs (#All negative ROIs) are listed for the training, validation, and test sets. The number of positive, labeled negative, all negative ROIs are computed using Sv threshold at -66 dB re 1 m^{-1}. With different Sv settings, these numbers change accordingly.

Split	#Echograms	#Annotated echograms	#Labeled positive ROIs	#Labeled negative ROIs	#All negative ROIs
Training	715	435	715	260	15,948
Validation	239	149	193	114	5873
Test	238	143	215	45	5188

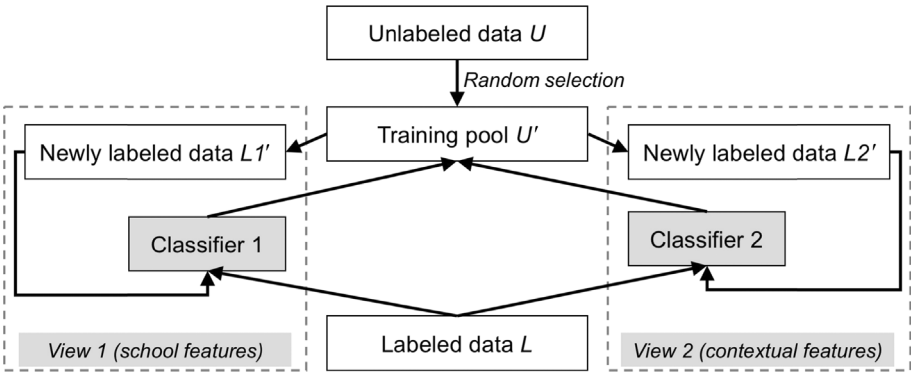

Fig. 5. Multiview co-training algorithm for ROI classification. A training pool (U') is randomly selected from the unlabeled data (U). Classifier 1 is then trained on labeled data (L) and any newly labeled data ($L1'$) using school features. Classifier 2 is trained on labeled data (L) and newly labeled data ($L2'$) using contextual features. Subsets of unlabeled data are fed back into the U' pool.

(Zheng 2015). As shown in "Annotation Characteristics" section, for Atlantic herring schools, their depths were related to the school characteristics, frequency responses, and aspect ratio. Based on our multiview experimental results shown in "Multiview-Based Model Performance" section, ROI classification performance was not degraded by this relationship.

Figure 5 and Algorithm 1 illustrate the process of co-training. In addition to labeled data L and unlabeled data U, a training pool U' was created by randomly choosing u' samples from U. During each iteration, Classifiers 1 and 2 were trained using L and then used to label m samples from U' with high confidence (m included p-positive and n-negative samples, which strictly followed the original ratio of positive to negative labels). With each iteration, 2-m samples were labeled. Once the samples in U' were labeled as either positive or negative, they were removed from U' and added into L. This process was conducted in an iterative manner until the stopping criteria, namely no unlabeled data were left or the maximum number of iterations (K in Algorithm 1) was reached. Therefore, K can be used to control the size of unlabeled data. When $K = 1$, $C1$ and $C2$ were trained only using labeled data L. The co-training algorithm was implemented with the default parameters used in the original paper (Blum and Mitchell 1998). When using the co-training model for inference, a label was assigned either when the two classifiers agreed with each other or by the classifier with higher confidence.

Multiview learning features

Depending on the approach to generate school-related features, there were two options for Classifier 1 as shown in Figure 5: (1) train a shallow machine learning model using the HC features and (2) train a CNN classifier. For each option, the machine learning and CNN model with the best performance were selected to be used in the co-training algorithm. For Classifier 2, machine learning models were trained using contextual features.

HC features: As shown in Table 3, a wide variety of acoustic and geometric features were extracted from the ROIs. Acoustic features include minimum, maximum, percentiles (5%, 25%, 50%, 75%, 95%), and standard deviation of each frequency component's Sv values, and the Sv values at 18, 120, and 200 kHz relative to 38 kHz. Geometric parameters that have been previously used in acoustic target classification (Haralabous and Georgakarakos 1996; Reid 2000; Korneliussen et al. 2009) were also computed, including length, thickness, perimeter, area, circularity, image compactness, elongation, and rectangularity. The basic unit of length and thickness (as illustrated in Fig. 7) was meters rather than pixels, which takes into account varying vessel speeds and provides a more accurate geometric characterization of the fish or krill schools. The HC features presented in Table 3 cover a comprehensive set of school-related features. Using these features, three different machine learning models (random forest [RF], support vector machine [SVM], and logistic regression [LR]) were trained to conduct ROI classification.

CNN features: To prepare inputs for the CNN models, all the ROIs were first converted into 4-channel images corresponding to the 4 frequencies: 18, 38, 120, and 200 kHz by cropping the bounding boxes of ROIs from echograms and resizing them to 100 × 100. It is worth noting that this process can cause object distortion, but transforming into fix-sized images is a requirement of the CNN models. The Sv values were scaled to the range of 0–255. Figure 6 shows the ResNet18 architecture for ROI classification. ResNet (He et al. 2016) leverages a novel architecture called "skip connection" that helps to alleviate the problem of gradient vanishing and has achieved state-of-the-art accuracy in many tasks. In addition to ResNet18, two other CNN architectures were explored: VGG16 (Simonyan and Zisserman 2014) and MobileNet v2 (Howard et al. 2017). VGG16 consists of 16 convolutional layers and has a uniform architecture, compared to ResNet18, which has more parameters and higher complexity. MobileNet v2 makes use of depth-wise separable convolutions, which significantly reduces the number of parameters.

Compared with ResNet18, MobileNet v2 is lighter and faster. For model training, batch size was set to 128, Adam optimizer was used with a learning rate of 0.001, and each model was trained for 100 epochs using the Cross-Entropy

Table 3. Detailed descriptions of the HC features. "*" can be replaced by different statistics listed in the description column, for example, Sv 38 kHz min. "std" denotes standard deviation.

Category	Feature	Unit	Description
Acoustic	Sv_18kHz_*	dB re 1 m^{-1}	Min, max, percentiles, std
	Sv_38kHz_*	dB re 1 m^{-1}	Min, max, percentiles, std
	Sv_120kHz_*	dB re 1 m^{-1}	Min, max, percentiles, std
	Sv_200kHz_*	dB re 1 m^{-1}	Min, max, percentiles, std
	Sv_relative_*		Sv at 18, 120, and 200 kHz relative to 38 kHz
Geometric	Length	m	Length of the bounding box
	Thickness	m	Width of the bounding box
	Perimeter	m	Perimeter of the fish school
	Area	m^2	Area of the fish school
	Circularity		$4*\pi*\text{area}/\text{perimeter}^2$
	Image compactness	m^{-1}	Perimeter/area
	Elongation		Length/thickness
	Rectangularity		(Length × thickness)/area

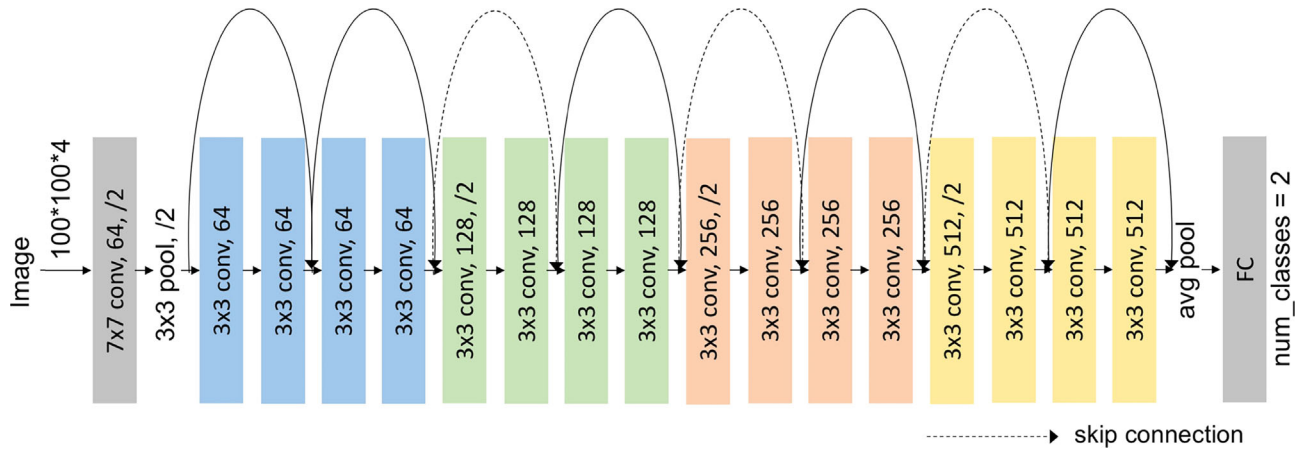

Fig. 6. ResNet18 architecture.

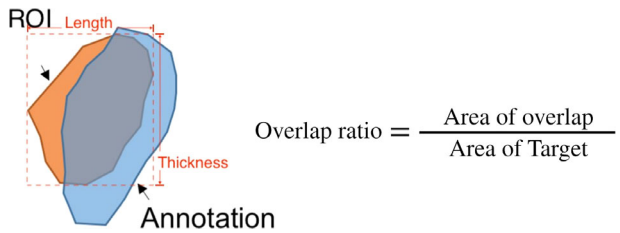

Fig. 7. Illustration of how the overlap ratio is calculated for the ROI and annotation or "target." School length and thickness described in Table 3 are annotated as well.

loss. The models were implemented with PyTorch[4] and Skorch,[5] and trained on a NVIDIA GeForce GTX 1080 GPU.

[4]https://github.com/pytorch/pytorch

[5]https://github.com/skorch-dev/skorch

Contextual features: Contextual features were extracted from information describing the vertical and horizontal geolocation of ROIs. There are five different features for each ROI: depth, total water column, relative altitude (ratio of depth to total water column), latitude, and longitude. These features were used in the RF, SVM, and LR machine learning models trained to conduct ROI classification.

Evaluation

The whole framework was evaluated using the test set. The hybrid model was designed to capture as many Atlantic herring schools as possible and be able to discriminate Atlantic herring schools from other classes of objects. Therefore, recall, precision, and *F*1-score were used to evaluate its overall performance (Davis and Goadrich 2006). These metrics have been widely used to evaluate the performance of classification models (Goutte and Gaussier 2005).

Algorithm 1. Co-training algorithm for ROI classification

Input: L—a set of labeled samples, U—a set of unlabeled samples

function co-training (L, U)
  Create a pool U' by randomly choosing u' samples from U
  $k \leftarrow 0$
  while U is not empty and $k < K$ **do** ▷ K: maximum iteration
    Train $C1$ using L ▷ $C1$: ROI Classifier 1
      Option 1: train machine learning model with handcrafted features
      Option 2: train convolutional neural networks
    Train $C2$ using L ▷ $C2$: ROI Classifier 2
    Let $C1$ label m samples from U' with most confident predictions ▷ m: p positive and n negative samples
    Let $C2$ label m samples from U' with most confident predictions
    Remove 2 m samples from U' and add them into L
    Randomly choose 2 m samples from U to replenish U'
    $k \leftarrow k + 1$
  end while
end function

Recall was used to evaluate the effectiveness of the framework in detecting Atlantic herring schools. It was computed as the fraction of annotations of Atlantic herring schools that were successfully captured by the framework to the number of true positives and number of false negatives. In other words, it is the ratio of the number of correctly classified Atlantic herring schools to the summation of the number of correctly classified Atlantic herring schools and the number of Atlantic herring schools incorrectly classified as something else.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (1)$$

Recall ranges from 0 to 1, and a higher value indicates that more Atlantic herring schools are correctly identified by the framework. Recall accounts for the number of non-Atlantic herring schools that were falsely classified as Atlantic herring (a Type II error), is analogous to statistical accuracy where higher values indicate better classification of the "true" Atlantic herring schools and can be a better evaluation metric for unbalanced distribution of classes. Recall is sensitive to false negatives and is useful if the "cost," for example, error in abundance estimates, of not correctly classifying Atlantic herring is significant. In our case, recall is quite important to estimating accurate abundance.

To claim a successful capture, there must be a significant portion of the annotation being covered. As shown in Figure 7, by overlapping ROIs that were predicted as Atlantic herring schools with annotations of Atlantic herring schools, an overlap ratio was computed for each pair. Here the "target" is the annotation. When the minimum overlap ratio is set to 0.5, it means that there must be at least 50% of the annotation being covered by the ROI. Varying minimum overlap ratios (0.3, 0.5) were experimented with, and higher ratio requires a larger overlap with annotations. Recall is analogous to the statistical term of accuracy, where recall is how well the algorithm selected the correct (i.e., "true") label. In this case, the annotations were assumed to be the "truth."

Precision was used to evaluate the effectiveness of the model at classifying Atlantic herring schools. Precision is the ratio of true positives to the summation of true and false positives (Type I errors). In other words, it is the ratio of the number of correctly classified Atlantic herring schools to the number of correctly classified Atlantic herring schools and the number of falsely classified schools as Atlantic herring.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

Precision ranges from 0 to 1, where a higher value indicates the model is better at classifying Atlantic herring schools. Precision is sensitive to false positives and is a useful metric if the "cost," for example, error in abundance estimates, is highly dependent on whether non-Atlantic herring backscatter was classified as Atlantic herring. For example, classifying other Clupeid species as Atlantic herring could result in significantly biased abundance estimates when their abundance is high.

$F1$-score is the weighted average of recall and precision, which ranges from 0 to 1 and was computed as follows:

$$F_1\text{-score} = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}} \quad (3)$$

The $F1$-score is a balance of precision and recall and can be interpreted as the harmonic mean of precision and recall. It is a useful metric when information about false positives and negatives are important to understand. In the case of estimating abundance for a target species, recall is potentially more

informative for evaluating the effectiveness of the models than precision, except in the case where confounding species are abundant, and the *F*1-score is a good general criterion.

Results

The performance of the proposed framework under different settings is presented in four parts. First, the hybrid model with single-view and multiview learning was tested and compared to evaluate the effectiveness of multiview learning. Second, supervised and semi-supervised models were tested and compared to evaluate the effectiveness of leveraging unlabeled data. Third, a parameter study was conducted to examine the impact of parameters on the model performance, such as Sv threshold and ratio of negative to positive samples. Fourth, example results of the proposed framework to identify Atlantic herring schools from other objects in acoustic data are provided.

Single-view vs. multiview learning

The ROI classification can be conducted either with single-view (only one feature set) or multiview learning. To perform a fair comparison between single-view- and multiview-based models, unlabeled data were not used at this stage. This approach separates the effect of multiview learning and unlabeled data. By setting iteration number K as 1, Classifiers 1 and 2 ($C1$ and $C2$) as described in Figure 5 were trained only on the labeled data. In "Supervised vs. Semi-supervised Learning" section, separate experiments were conducted to evaluate the effectiveness of unlabeled data.

Single-view-based model performance

Tables 4 and 5 provide the recall, precision, and *F*1-scores of single-view-based models at minimum overlap ratio of 0.3 and 0.5, respectively. With a minimum overlap ratio of 0.5, the recall and *F*1-scores of all models decreased. When testing on *positive and labeled negative samples*, $C2$ with contextual features achieved the best performance with an *F*1-score of 0.850 and 0.889 at minimum overlap ratio of 0.5 and 0.3, respectively.

The finding that inclusion of contextual features is the most effective aligns with the observation that school depth is an important feature for separating Atlantic herring from other Clupeid species. However, the results are different when testing on *positive and all negative samples*. In this scenario, $C2$ with contextual features had the lowest *F*1-score of 0.445 and 0.460, and $C1$ with HC features achieved the best performance with an *F*1-score of 0.666 and 0.707. All negative samples included negative samples outside the labeled nontarget classes, and the results here indicate that using only contextual features is not sufficient to differentiate Atlantic herring schools from all other nontarget classes. Also, when switching from *labeled negatives* to *all negatives*, all models showed some decrease in precision and *F*1-score. $C1$ with CNN features and $C2$ with contextual features had a larger performance drop than $C1$ with HC features, which demonstrated that the HC features are more robust when applying to more diverse cases. For models using the same feature set, RF achieved the best performance with HC features and with contextual features while ResNet18 achieved the best performance among CNN models. These best-performing methods were used in the following multiview learning experiments.

Another finding from Tables 4 and 5 is that $C1$ with HC features outperformed $C1$ with CNN features. There are several reasons for this: (i) the size of labeled data is relatively small, as shown in Table 2, and the number of positive samples for training is less than 1,000. A previous study similarly showed that when the size of labeled data is small, the HC features outperformed CNN features (Lin et al. 2020). (ii) When transforming acoustic data into fixed size images, the objects may be distorted and cause performance degradation in the CNN models. Distortion is a natural problem in acoustic data due to the impact of varying vessel speed, differences in scale where the vertical extent is much larger than the horizontal extent, and the difficulty keeping accurate geometric information with only the pixel-based image. Figure 8 shows the top 20 important features of the RF using HC features. The most important features are thickness and compact (both are

Table 4. Performance of single-view-based models using HC, CNN, or contextual features with a minimum overlap ratio of 0.3.

		Positives and labeled negatives			Positives and all negatives		
Classifier	**Model**	**Recall**	**Precision**	***F*1-score**	**Recall**	**Precision**	***F*1-score**
$C1$ (HC)	RF	0.671	0.892	0.766	0.671	0.746	0.707
	SVM	0.597	0.870	0.708	0.597	0.596	0.596
	LR	0.326	0.824	0.467	0.326	0.479	0.388
$C1$ (CNN)	ResNet	0.637	0.842	0.725	0.637	0.546	0.588
	MobileNet	0.573	0.814	0.672	0.573	0.476	0.520
	VGG	0.573	0.802	0.668	0.573	0.429	0.490
$C2$ (contextual)	RF	0.803	0.947	0.869	0.803	0.322	0.460
	SVM	0.824	0.965	0.889	0.824	0.264	0.400
	LR	0.318	0.778	0.452	0.318	0.172	0.223

Table 5. Performance of single-view-based models using HC, CNN, or contextual features with a minimum overlap ratio of 0.5.

		Positives and labeled negatives			Positives and all negatives		
Classifier	**Model**	**Recall**	**Precision**	***F*1-score**	**Recall**	**Precision**	***F*1-score**
C1 (HC)	RF	0.601	0.892	0.718	0.601	0.746	0.666
	SVM	0.541	0.870	0.668	0.541	0.596	0.567
	LR	0.275	0.824	0.413	0.275	0.479	0.350
C1 (CNN)	ResNet	0.563	0.842	0.675	0.563	0.546	0.554
	MobileNet	0.515	0.814	0.631	0.515	0.476	0.495
	VGG	0.510	0.802	0.624	0.510	0.429	0.466
C2 (contextual)	RF	0.717	0.947	0.816	0.717	0.322	0.445
	SVM	0.760	0.965	0.850	0.760	0.264	0.392
	LR	0.285	0.778	0.417	0.285	0.172	0.215

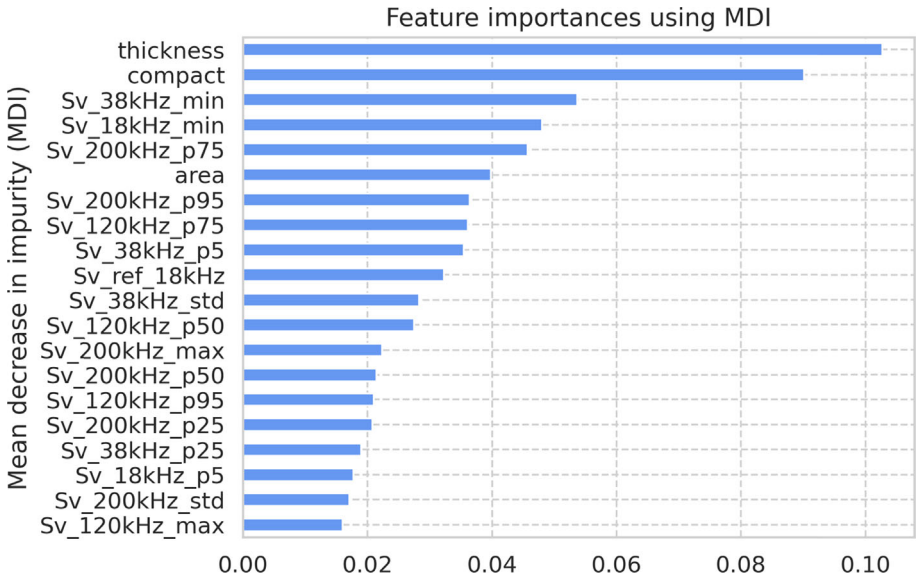

Fig. 8. Top 20 important features of the RF model with HC features. Thickness, compact, and area are geometric features while the remaining variables derive from the acoustic data. See Table 3 for further details on the feature descriptions. Mean decrease in impurity (MDI) is an impurity-based feature importance.

geometric features), which demonstrate the importance of accurate geometric features for ROI classification.

Multiview-based model performance

Tables 6 and 7 provide the recall, precision, and *F*1-score of multiview-based models at minimum overlap ratios of 0.3 and 0.5, respectively. When testing on *positives and labeled negatives*, co-training models performed worse than *C*2 only but better than *C*1 only. The results indicated that multiview learning was not beneficial in differentiating Atlantic herring schools from the labeled nontarget classes. It is likely that *C*2 already has sufficient discriminating ability and combining it with *C*1 may degrade its performance. For this scenario, using *C*2 with contextual features was better. When testing on *positives and all negatives*, the results were different. By combining *C*1 with HC features and *C*2 with contextual features, co-training models outperformed *C*1 and *C*2 only and achieved an *F*1-score of 0.758 and 0.804 at minimum overlap ratios of 0.5 and 0.3, respectively. The results indicated that multiview learning was beneficial in differentiating Atlantic herring schools from all other nontarget classes. However, this was not true when combining *C*1 with CNN features and *C*2 with contextual features, that is, the co-training models performed better than *C*2 only but worse than *C*1 only. An important assumption behind the co-training algorithm was that the two views/classifiers were providing complementary information to each other. The results showed HC and contextual features did benefit each other, while CNN and contextual features did not. In addition, co-training models using HC features outperformed the ones using CNN features, which is also consistent with single-view model results.

Supervised vs. semi-supervised learning

Besides multiview learning, another important advantage of the co-training algorithm was its utilization of unlabeled data. To perform a fair comparison between supervised and semi-supervised methods, co-training models with and without unlabeled data were tested and compared. Both models used $C1$ with HC features and $C2$ with contextual features. For the co-training models with unlabeled data, maximum iteration K was set to 15, which corresponded to 210 unlabeled data since the positive: negative ratio was set to 1:6 (a detailed study on the optimal positive: negative ratio is provided in "ROI Classification: Ratio of Negative to Positive Samples" section).

As shown in Figure 9, the co-training models were trained using a varying number of labeled data. The number of positive data started from 17, with an increment step of 7, up to 357 (about 50% of the positive data shown in Table 2). When the number of positive data was below 100, the co-training model with unlabeled data consistently outperformed the model without unlabeled data, which demonstrated the effectiveness of utilizing unlabeled data when the training set was small. However, when the number of positive data increased to 100 or larger, there was no advantage and the co-training model with unlabeled data achieved F1-Scores comparable to the model without unlabeled data. These results indicated that the utilization of unlabeled data is useful when the training set was small.

Parameter study

The parameter study was conducted on the validation set. Key parameters like Sv threshold and ratio of negative to positive samples were selected to study their impact on model performance. The minimum overlap ratio was set to 0.3.

ROI detection: Sv threshold

Figure 10 provides the ROI detection and classification results at varying Sv thresholds. For ROI detection, the recall at -68 to -66 dB re 1 m^{-1} is nearly 1.0, which indicated that this Sv threshold detected almost all the Atlantic herring schools that were annotated. This range is also well aligned with the empirical value of -66 dB re 1 m^{-1}. For ROI classification, the highest recall and $F1$-score values were achieved at

Table 6. Performance of multiview-based models with a minimum overlap ratio of 0.3.

	Positives and labeled negatives			Positives and all negatives		
Classifier	**Recall**	**Precision**	***F*1-score**	**Recall**	**Precision**	***F*1-score**
Co-training (HC)	0.788	0.964	0.868	0.788	0.819	0.804
Co-training (CNN)	0.654	0.917	0.764	0.654	0.380	0.481

Table 7. Performance of multiview-based models with a minimum overlap ratio of 0.5.

	Positives and labeled negatives			Positives and all negatives		
Classifier	**Recall**	**Precision**	***F*1-score**	**Recall**	**Precision**	***F*1-score**
Co-training (HC)	0.705	0.964	**0.814**	0.705	0.819	**0.758**
Co-training (CNN)	0.589	0.917	0.718	0.589	0.380	0.462

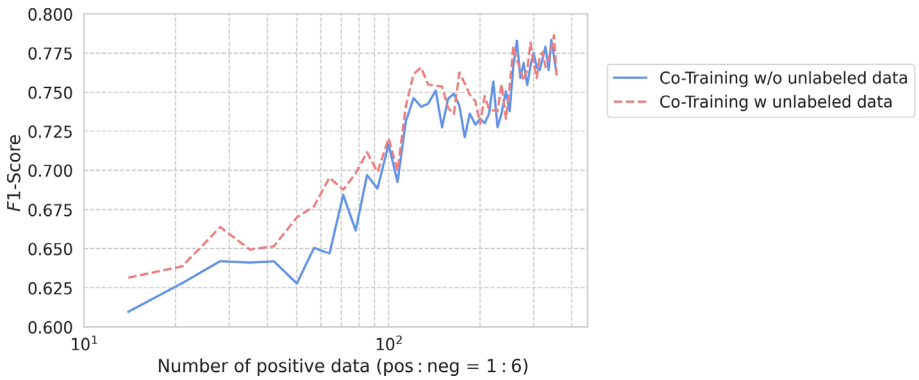

Fig. 9. Performance of supervised (blue line) and semi-supervised learning (red dashed line) with a minimum overlap ratio of 0.3.

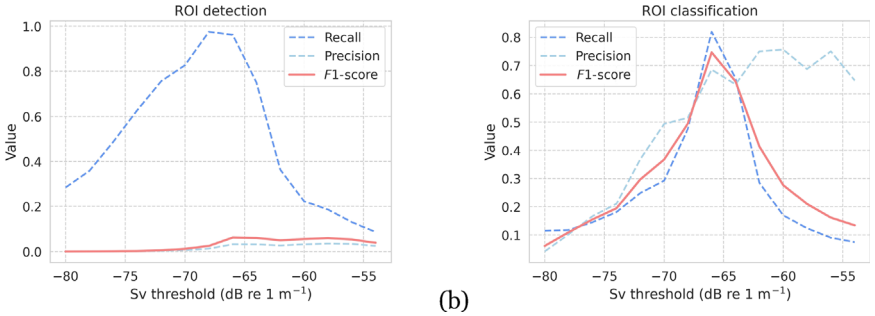

Fig. 10. ROI detection and classification performances on the validation set as Sv threshold varied between −80 and −54 dB re 1 m^{-1} with an interval of 2 dB re 1 m^{-1}.

−66 dB re 1 m^{-1}, which was similar to the optimal Sv threshold in the ROI detection. Additionally, the precision increased significantly after the ROI classification step, which suggested that the classification step was essential in discriminating Atlantic herring schools from other classes. Given the results, the optimal Sv threshold was selected as −66 dB re 1 m^{-1}.

ROI classification: Ratio of negative to positive samples

After setting the Sv threshold to −66 dB re 1 m^{-1}, the number of positive and negative samples are presented in Table 2. Figure 11 provides the ROI classification results at varying ratios of negative to positive samples (from 1 to 10). Higher ratios mean more negative samples, and the model was more likely to classify a sample as negative. Overall, the recall decreased and the precision increased as the ratio increased. According to the *F*1-score results, the optimal ratio was 6 (positive: negative = 1:6) which best balanced recall and precision.

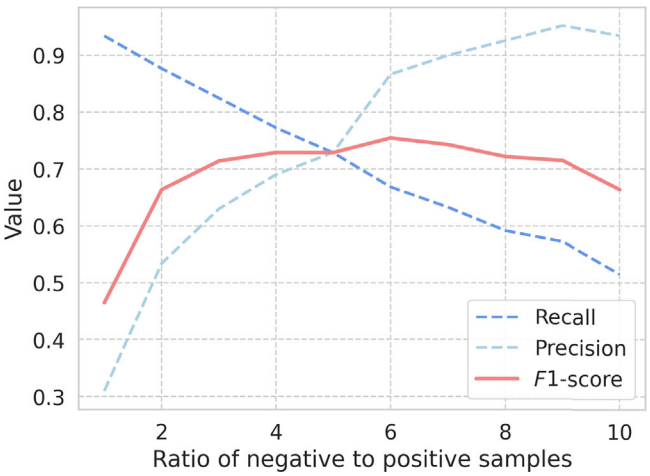

Fig. 11. ROI classification performance on the validation set as the ratio of positive to negative samples varied from 1 to 10.

Prediction examples

Figure 12 provides a comparison between annotations and predictions made by applying the proposed framework on test examples. In example A, some annotations of "Atlantic herring school" were missing at depth of 100–150 m. The framework successfully identified all the Atlantic herring schools including those in the missing regions, that is, those missed by the human scrutinizer. However, some seabed echos were misclassified as Atlantic herring schools. In Example B, most Atlantic herring schools appeared near the seafloor, and the predictions were consistent with the annotations. In Example C, there were annotations of "Noise region" near the sea surface, and annotations of "Atlantic herring school" near the seafloor. The framework correctly classified both cases. However, in the region at 150 m depth, there were no annotations but the framework identified some Atlantic herring schools which could be some misclassifications. In Example D, there were some annotations of "Clupeid school," and the framework correctly classified them as not Atlantic herring schools.

Discussion

We explored multiview learning to develop a hybrid model for Atlantic herring school detection. The hybrid model consisted of two components: ROI detection and ROI classification. For ROI detection, a contour detection-based method was applied to detect Atlantic herring schools and other objects. We allowed the Sv threshold to vary between −80 and −54 dB re 1 m^{-1} in the training process, where it was included as a hyperparameter of the framework, and the appropriate value was chosen by optimizing the *F*1-score on the validation set. The model's finding that an Sv threshold of −66 dB showed the best performance was consistent with the historical Sv threshold used by the NEFSC (Jech and Michaels 2006) and provided good confidence in our approach as this threshold is nearly universally employed by fisheries institutions that use acoustic data for assessment of herring as well as a few other gas-bearing species. For ROI classification, the co-training algorithm was employed because it enables both multiview learning and utilization of unlabeled data. Comparisons were conducted between single-view and

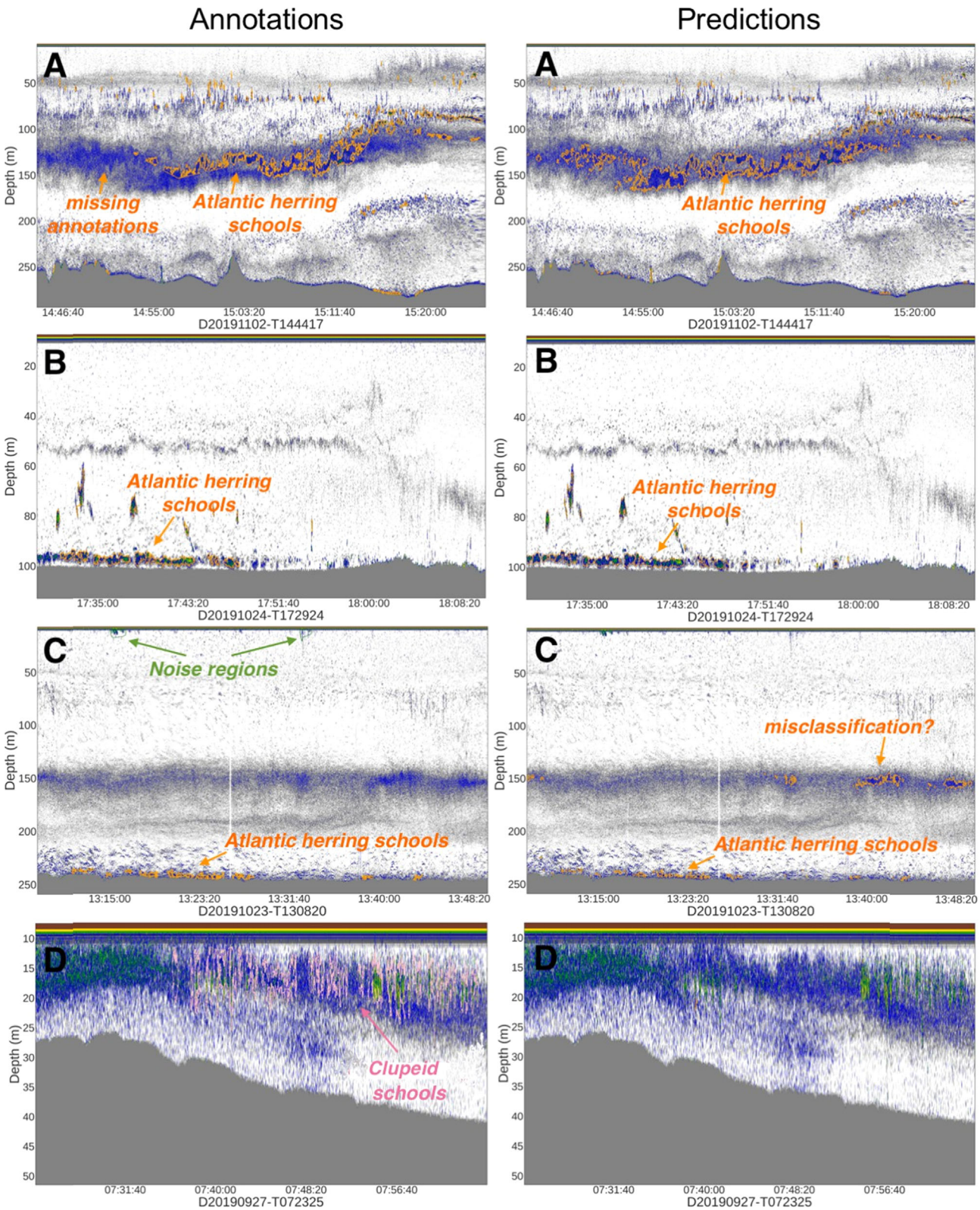

Fig. 12. Comparison of (left) annotations and (right) predictions. The colors denote different annotation classes (orange: "Atlantic herring school," pink: "Clupeid school," green: "Noise region"). The date and start time are given at the bottom of each echogram, for example, D20190927-T072325 represents data collected on 27 September 2019 starting at 07:23:25 UTC. The gray portions of the echograms are below the seabed detection. Echograms show 38-kHz Sv with a color scale of −80 to 0 dB re 1 m^{-1} from gray to blue to green. Sv values less than −80 dB re 1 m^{-1} have been excluded and shown as the white background.

multiview-based models, as well as supervised and semi-supervised learning models.

The co-training model with HC and contextual features outperformed other models in distinguishing Atlantic herring schools from other nontarget objects. It was also shown that the utilization of unlabeled data improves the model performance when the training set is small. The detection and classification of Atlantic herring schools by the proposed framework matched well with those by experts who have many years of experience, but with subjectivity and constrained time. These results are important for maintaining a consistent time series such as annual abundance estimates included in marine resource assessments; for programs that are resource limited and/or do not have multiple personnel with manual scrutiny expertise; and reducing or eliminating subjective bias by providing an explicit work flow that can be used and revised as methods improve.

As such, the framework proposed in this work is a step towards building a fully automatic system for Atlantic herring school detection that is accurate, time-effective, and significantly reduces manual efforts for the NEFSC. For example, the NEFSC fisheries acoustics expert took on average 2 h to generate annotations for each day of data, and in total about 80 h to generate all annotations for the 2019 acoustic data. In contrast, the proposed framework only took about 5 h to process the same amount of data. Additionally, this new framework was implemented in the open-source computing language Python and built upon publicly available PyEcholab. As larger amounts of acoustic data are archived and made publicly available (Wall et al. 2016), this kind of automatic system is highly desirable, especially those implemented using open-source software.[6]

The proposed framework enables the analysis of both historical and future surveys of Atlantic herring. For historical surveys, this framework can be integrated with the existing procedure of Atlantic herring abundance estimation. For future surveys, this framework could be used to quickly identify Atlantic herring schools in real time.

This study aligns well with many recent studies employing machine learning and deep learning techniques in acoustic target identification. Previous works have used artificial neural networks (Haralabous and Georgakarakos 1996; Reid 2000; Cabreira et al. 2009), RF (Fallon et al. 2016; Proud et al. 2020), CNN-based models like ResNet (Rezvanifar et al. 2019), U-Net (Brautaset et al. 2020), and YOLO (Porto Marques et al. 2021) in acoustic target classification or detection. A major benefit of CNN-based models is their capability in automatic feature extraction; however, those features are usually limited to images. To boost the performance of CNN-based models, there have been studies exploring new model structures that can incorporate contextual information in classification (Chu and Cai 2018; Ellen et al. 2019; Hu et al. 2019). To the best of our knowledge, contextual information is still not well used in the object detection model. This study is built upon the hybrid model, which has a more flexible structure compared to the end-to-end models, and can incorporate contextual information via multiview learning. While finding the optimal object detection model is outside the scope of this study, we plan to explore approaches to incorporate contextual information in the end-to-end models as well and perform a comprehensive comparison of the hybrid and end-to-end models in future work.

By utilizing unlabeled data in the co-training algorithm, this study examines semi-supervised learning in the ROI classification. Semi-supervised learning methods have been shown to be effective when there is a small number of annotations (Zhu and Goldberg 2009). The scarcity of labeled data is a common issue in acoustic target classification/detection since the annotations are generated with manual scrutiny and require a significant investment in time and expertise by the human scrutinizer. A recent study by Choi et al. (2021) showed the semi-supervised learning method outperformed supervised learning methods in the task of sandeel classification by providing better accuracy with fewer annotations. The results from this study further demonstrate this point and showed that unlabeled data can help improve model performance when the labeled dataset is small.

The performance of acoustic target detection depends on the species, school-related features, object detection method, and the annotations. As shown in Fig. 12, there were still some discrepancies between annotations and predictions. These discrepancies were mainly caused by missing annotations or erroneous predictions. More accurate predictions can be achieved by (i) preparing a set of annotations with confidence scores where the scores are added to the annotations and less or uncertain annotations can be excluded from or given less weight during model training; (ii) increasing the variability of annotations in order to include cases that span the full variability of the target species, such as Atlantic herring schools at different depths; and (iii) adding other negative classes of annotations, as shown in Fig. 12, some seabed echos were misclassified as Atlantic herring schools. If annotations for these scenarios are leveraged by the ROI classifier, classification accuracy can be further improved.

There are limitations of this study, which can direct future directions to explore. In this study, the framework was trained and tested using data collected during a single-year survey in a specific region. In our case where the assessment surveys are conducted at the same time of year and in the same geographic region, we anticipate that these results should be applicable to other years. However, as the waters in the Gulf of Maine warm (Pershing et al. 2015, 2021), the timing, locations, and behavior of spawning Atlantic herring could change, which may affect the generality of our model. In addition, the optimal Sv threshold and ratio of negative to positive

[6]https://github.com/ices-eg/wg_WGFAST/tree/master/Open-Source_Effort

samples that were optimized with the validation set may not be directly applicable to other regions. As a future direction, we plan to extend the framework to train and test with multi-year survey data to evaluate its generality across years. Also, some features used in this study are region-specific, such as latitude and longitude. When applying the framework to different regions, it is worth exploring features that are not region-specific but still effective in representing geographic preference of the target species, such as environmental factors like temperature and productivity (Escobar-Flores et al. 2013).

Data availability statement

The acoustic data used in this study are archived at the NOAA National Centers for Environmental Information (NCEI) (2020; NOAA Northeast Fisheries Science Center 2019) and publicly accessible for free through the archive's data portal at https://www.ncei.noaa.gov/maps/water-column-sonar/ and Amazon Web Services (AWS) at https://registry.opendata.aws/ncei-wcsd-archive/. The code repository of this study is https://github.com/yawenzzzz/AH-school-detection.

References

Arbelaez, P., M. Maire, C. Fowlkes, and J. Malik. 2010. Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**: 898–916. doi:10.1109/TPAMI.2010.161

Barange, M. 1994. Acoustic identification, classification and structure of biological patchiness on the edge of the Agulhas Bank and its relation to frontal features. S. Afr. J. Mar. Sci. **14**: 333–347. doi:10.2989/025776194784286969

Beyan, C., and H. I. Browman. 2020. Setting the stage for the machine intelligence era in marine science. ICES J. Mar. Sci. **77**: 1267–1273. doi:10.1093/icesjms/fsaa084

Blum, A., and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training, p. 92–100. *In* Proceedings of the Eleventh Annual Conference on Computational Learning Theory. ACM. doi:10.1145/279943.279962

Brandt, S., and D. McEvoy. 2006. Distributional effects of property rights: Transitions in the Atlantic herring fishery. Mar. Policy **30**: 659–670. doi:10.1016/j.marpol.2005.09.007

Brautaset, O., A. U. Waldeland, E. Johnsen, K. Malde, L. Eikvil, A.-B. Salberg, and N. O. Handegard. 2020. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. ICES J. Mar. Sci. **77**: 1391–1400. doi:10.1093/icesjms/fsz235

Cabreira, A. G., M. Tripode, and A. Madirolas. 2009. Artificial neural networks for fish-species identification. ICES J. Mar. Sci. **66**: 1119–1129. doi:10.1093/icesjms/fsp009

Choi, C., M. Kampffmeyer, N. O. Handegard, A.-B. Salberg, O. Brautaset, L. Eikvil, and R. Jenssen. 2021. Semi-supervised target classification in multi-frequency echosounder data. ICES J. Mar. Sci. **78**: 2615–2627. doi:10.1093/icesjms/fsab140

Chu, W., and D. Cai. 2018. Deep feature based contextual model for object detection. Neurocomputing **275**: 1035–1042. doi:10.1016/j.neucom.2017.09.048

Coetzee, J. 2000. Use of a shoal analysis and patch estimation system (SHAPES) to characterise sardine schools. Aquat. Living Resour. **13**: 1–10. doi:10.1016/S0990-7440(00)00139-X

Davis, J., and M. Goadrich. 2006. The relationship between precision-recall and ROC curves, p. 233–240. *In* Proceedings of the 23rd International Conference on Machine Learning. ACM. doi:10.1145/1143844.1143874

Ellen, J. S., C. A. Graff, and M. D. Ohman. 2019. Improving plankton image classification using context metadata. Limnol. Oceanogr. Methods **17**: 439–461. doi:10.1002/lom3.10324

Escobar-Flores, P., R. L. O'Driscoll, and J. C. Montgomery. 2013. Acoustic characterization of pelagic fish distribution across the south Pacific Ocean. Mar. Ecol. Prog. Ser. **490**: 169–183. doi:10.3354/meps10435

Fallon, N. G., S. Fielding, and P. G. Fernandes. 2016. Classification of Southern Ocean krill and icefish echoes using random forests. ICES J. Mar. Sci. **73**: 1998–2008. doi:10.1093/icesjms/fsw057

Foote, K. G. 1987. Calibration of acoustic instruments for fish density estimation: A practical guide, p. 1–69. *In* ICES Cooperative Research Reports, v. **144**. ICES. doi:10.17895/ices.pub.8265

Girshick, R. 2015. Fast R-CNN, p. 1440–1448. *In* Proceedings of the IEEE International Conference on Computer Vision. IEEE. doi:10.1109/ICCV.2015.169

Goutte, C., and E. Gaussier. 2005. A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation, p. 345–359. *In* European Conference on Information Retrieval. Springer. doi:10.1007/978-3-540-31865-1_25

Haralabous, J., and S. Georgakarakos. 1996. Artificial neural networks as a tool for species identification of fish schools. ICES J. Mar. Sci. **53**: 173–180. doi:10.1006/jmsc.1996.0019

Hastie, T., R. Tibshirani, and J. Friedman. 2009. Overview of supervised learning, p. 9–41. *In* The elements of statistical learning. Springer. doi:10.1007/978-0-387-84858-7_2

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition, p. 770–778. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. doi:10.1109/CVPR.2016.90

Horne, J. K. 2000. Acoustic approaches to remote species identification: A review. Fish. Oceanogr. **9**: 356–371. doi:10.1046/j.1365-2419.2000.00143.x

Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. doi:10.48550/arXiv.1704.04861

Hu, S., R. Ma, and H. Wang. 2019. An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. PloS One **14**: e0225317. doi:10.1371/journal.pone.0225317

Jech, J. M., W. Michaels, W. Overholtz, W. Gabriel, T. Azarovitz, D. Ma, K. Dwyer, and R. Yetter. 2000. Fisheries acoustic surveys in the Gulf of Maine and on Georges Bank at the Northeast Fisheries Science Center, p. 1–3. *In* Proceedings of the Sixth International Conference on Remote Sensing for Marine and Coastal Environments. Veridian ERIM International.

Jech, J. M., and W. L. Michaels. 2006. A multifrequency method to classify and evaluate fisheries acoustics data. Can. J. Fish. Aquat. Sci. **63**: 2225–2235. doi:10.1139/f06-126

Jech, J. M., and F. Stroman. 2012. Aggregative patterns of prespawning Atlantic herring on Georges Bank from 1999–2010. Aquat. Living Resour. **25**: 1–14. doi:10.1051/alr/2012003

Jech, J. M., and P. J. Sullivan. 2014. Distribution of Atlantic herring (*Clupea harengus*) in the Gulf of Maine from 1998 to 2012. Fish. Res. **156**: 26–33. doi:10.1016/j.fishres.2014.04.016

Jech, J. M., G. Lawson, and M. Lowe. 2018. Comparing acoustic classification methods to estimate krill biomass in the Georges Bank region from 1999 to 2012. Limnol. Oceanogr. Methods **16**: 680–695. doi:10.1002/lom3.10275

Jørstad, K., D. King, and G. Nævdal. 1991. Population structure of Atlantic herring, *Clupea harengus* L. J. Fish Biol. **39**: 43–52. doi:10.1111/j.1095-8649.1991.tb05066.x

R. J. Korneliussen, Acoustic target classification 2018. ICES. doi:10.17895/ices.pub.4567

Korneliussen, R. J., Y. Heggelund, I. K. Eliassen, and G. O. Johansen. 2009. Acoustic species identification of schooling fish. ICES J. Mar. Sci. **66**: 1111–1118. doi:10.1093/icesjms/fsp119

Korneliussen, R. J., Y. Heggelund, G. J. Macaulay, D. Patel, E. Johnsen, and I. K. Eliassen. 2016. Acoustic identification of marine species using a feature library. Methods Oceanogr. **17**: 187–205. doi:10.1016/j.mio.2016.09.002

Li, H., and others. 2023. A multi-view co-training network for semi-supervised medical image-based prognostic prediction. Neural Netw. **164**: 455–463. doi:10.1016/j.neunet.2023.04.030

Lin, W., K. Hasenstab, G. Moura Cunha, and A. Schwartzman. 2020. Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. Sci. Rep. **10**: 1–11. doi:10.1038/s41598-020-77264-y

Malde, K., N. O. Handegard, L. Eikvil, and A.-B. Salberg. 2020. Machine intelligence and the data-driven future of marine science. ICES J. Mar. Sci. **77**: 1274–1285. doi:10.1093/icesjms/fsz057

NOAA National Centers for Environmental Information. 2020. Water column sonar data collection. NOAA National Centers for Environmental Information. doi:10.7289/V5HT2M7C

NOAA National Geophysical Data Center. 2009. Etopo1 1 arc-minute global relief model. NOAA National Centers for Environmental Information.

NOAA Northeast Fisheries Science Center. 2012 54th northeast regional stock assessment workshop (assessment summary report).

NOAA Northeast Fisheries Science Center. 2018. 65th northeast regional stock assessment workshop (assessment summary report).

NOAA Northeast Fisheries Science Center. 2019. Ek60 water column sonar data collected during hb1906. NOAA National Centers for Environmental Information. doi:10.25921/vt45-sa66

Pershing, A. J., and others. 2015. Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. Science **350**: 809–812. doi:10.1126/science.aac9819

Pershing, A. J., and others. 2021. Climate impacts on the Gulf of Maine ecosystem: A review of observed and expected changes in 2050 from rising temperatures. Elem. Sci. Anth. **9**: 00076. doi:10.1525/elementa.2020.00076

Politis, P. J., J. K. Galbraith, P. Kostovick, and R. W. Brown. 2014. Northeast fisheries science center bottom trawl survey protocols for the NOAA ship Henry B. Bigelow. *In* Northeast Fisheries Science Center Reference Document 2014, v. **14-06**. US Department of Commerce. doi:10.7289/V5C53HVS

Porto Marques, T., A. Rezvanifar, M. Cote, A. B. Albu, K. Ersahin, T. Mudge, and S. Gauthier. 2021. Detecting marine species in echograms via traditional, hybrid, and deep learning frameworks. *In* 2020 25th International Conference on Pattern Recognition (ICPR). IEEE. doi:10.1109/ICPR48806.2021.9412969

Proud, R., and others. 2020. Automated classification of schools of the silver cyprinid *Rastrineobola argentea* in Lake Victoria acoustic survey data using random forests. ICES J. Mar. Sci. **77**: 1379–1390. doi:10.1093/icesjms/fsaa052

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection, p. 779–788. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. doi:10.1109/CVPR.2016.91

D. G. Reid, Report on echo trace classification, 2000. ICES. doi:10.17895/ices.pub.5371

Reid, R., L. Cargnelli, S. Griesbach, D. Packer, D. Johnson, C. Zetlin, W. Morse, and P. Berrien. 1999. Atlantic herring, *Clupea harengus*, life history and habitat characteristics, p. 48. *In* NOAA Technical Memorandum NMFS-NE, v. **126**.

Rezvanifar, A., T. P. Marques, M. Cote, A. B. Albu, A. Slonimer, T. Tolhurst, K. Ersahin, T. Mudge, and S. Gauthier. 2019. A deep learning-based framework for the detection of schools of herring in echograms. *In* NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning,

Vancouver, Canada. arXiv preprint. doi:10.48550/arXiv.1910.08215

Shabangu, F. W., E. Ona, and D. Yemane. 2014. Measurements of acoustic attenuation at 38 kHz by wind-induced air bubbles with suggested correction factors for hull-mounted transducers. Fish. Res. **151**: 47–56. doi:10.1016/j.fishres.2013.12.008

Simonyan, K., and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. doi:10.1109/TPAMI.2014.2301163

Stephenson, R. L., G. D. Melvin, and M. J. Power. 2009. Population integrity and connectivity in northwest atlantic herring: A review of assumptions and evidence. ICES J. Mar. Sci. **66**: 1733–1739. doi:10.1093/icesjms/fsp189

Suzuki, S., and K. A. be. 1985. Topological structural analysis of digitized binary images by border following. Comput. Vis. Graph. Image Process. **30**: 32–46. doi:10.1016/0734-189X(85)90016-7

Trenkel, V. M., and L. Berger. 2013. A fisheries acoustic multi-frequency indicator to inform on large scale spatial patterns of aquatic pelagic ecosystems. Ecol. Indic. **30**: 72–79. doi:10.1016/j.ecolind.2013.02.006

Wall, C. C., J. M. Jech, and S. J. McLean. 2016. Increasing the accessibility of acoustic data through global access and imagery. ICES J. Mar. Sci. **73**: 2093–2103. doi:10.1093/icesjms/fsw014

Zhao, J., X. Xie, X. Xu, and S. Sun. 2017. Multi-view learning overview: Recent progress and new challenges. Inf. Fusion **38**: 43–54. doi:10.1016/j.inffus.2017.02.007

Zheng, Y. 2015. Methodologies for cross-domain data fusion: An overview. IEEE Trans Big Data **1**: 16–34. doi:10.1109/TBDATA.2015.2465959

Zhu, J., J. Shi, X. Liu, and X. Chen. 2014. Co-training based semi-supervised classification of Alzheimer's disease, p. 729–732. *In* 2014 19th International Conference on Digital Signal Processing. IEEE. doi:10.1109/ICDSP.2014.6900760

Zhu, X., and A. B. Goldberg. 2009. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning **3**: 1–130. doi:10.1007/978-3-031-01548-9

Acknowledgments

This research was supported by the University of Colorado Research and Innovation Seed Grant Program RISGP FY21. The NCEI Water Column Sonar Archive is supported by NOAA Fisheries through the NOAA cooperative agreement NA22OAR4320151, for the Cooperative Institute for Earth System Research and Data Science (CIESRDS). We would like to thank the Survey Technicians on the NOAA Ship HB Bigelow for recording and monitoring the EK60 echosounder data, and Chuck Anderson and Veronica Martinez for data archiving and dissemination.

Submitted 19 June 2023
Revised 24 January 2024
Accepted 01 March 2024

Associate editor: Paul F. Kemp