ICES Journal of Marine Science, 2024, Vol. 0, Issue 0, 1–14
https://doi.org/10.1093/icesjms/fsae020
Received: 14 October 2023; **revised:** 16 January 2024; **accepted:** 29 January 2024
Original Article

Fish age reading using deep learning methods for object-detection and segmentation

Arjay Cayetano [1,*], Christoph Stransky[1], Andreas Birk [2], Thomas Brey[3]

[1]Thünen Institute of Sea Fisheries, Bremerhaven 27572, Germany
[2]School of Science and Engineering, Constructor University, Bremen 28759, Germany
[3]Faculty of Biology and Chemistry, University of Bremen, Bremen 28334, Germany
*Corresponding author. Thünen Institute of Sea Fisheries, Bremerhaven 27572, Germany. E-mail: arjay.cayetano@thuenen.de

Abstract

Determination of individual age is one essential step in the accurate assessment of fish stocks. In non-tropical environments, the manual count of ring-like growth patterns in fish otoliths (ear stones) is the standard method. It relies on visual means and individual judgment and thus is subject to bias and interpretation errors. The use of automated pattern recognition based on machine learning may help to overcome this problem. Here, we employ two deep learning methods based on Convolutional Neural Networks (CNNs). The first approach utilizes the Mask R-CNN algorithm to perform object detection on the major otolith reading axes. The second approach employs the U-Net architecture to perform semantic segmentation on the otolith image in order to segregate the regions of interest. For both methods, we applied a simple postprocessing to count the rings on the output masks returned, which corresponds to the age prediction. Multiple benchmark tests indicate the promising performance of our implemented approaches, comparable to recently published methods based on classical image processing and traditional CNN implementation. Furthermore, our algorithms showed higher robustness compared to the existing methods, while also having the capacity to extrapolate missing age groups and to adapt to a new domain or data source.

Keywords: fish age reading; automation; deep learning; object detection; segmentation

Introduction

Individual age is an essential parameter in the analysis of fish population dynamics and thus a precondition for both sustainable management and a thorough understanding of the ecological role of a fish stock. The common approach in estimating the age of a fish is to make use of patterns along calcified structures such as scales and otoliths (ear stones) and observe the appearance of the annual growth zones (or annuli) (Panfili et al. 2002). These growth zones are formed by the uneven deposition of calcium carbonate and proteins as the fish experiences seasonal changes. Correspondingly, each single alternating opaque and translucent ring formation represents a period of one year (Campana 2001, Panfili et al. 2002). Hence, in traditional age reading, human experts perform manual counting of these ring patterns, which require individual judgment, especially if the rings are hardly distinguishable.

The pattern of ring formations can be distinct for each fish species, hence making the task of annual growth zone detection extremely challenging. Moreover, due to known environmental effects on otolith growth (Campana 1999), even different stocks of the same species can also have different ring patterns (Williams et al. 2005). In some cases, false rings and double rings can occur, which may lead to an overestimation of fish ages. Likewise, some rings can also be very faint and ambiguous, leading to underestimated age values (Campana 2001, Carbonara and Follesa 2019).

As otolith images and age data are collected in large quantities by various institutions as part of routine stock assessment, it is necessary to make the process of age reading scalable and less error-prone. In addition, the lack of age readers for a given species can also be a limitation due to the extensive nature of the training required. Even an expert on one species needs to be trained again for another species due to the differences in guidelines and protocols. Hence, it is not surprising that over the recent decades, a lot of attempts have been made to explore the possibility of automating the process. The first approaches were based on classical image processing techniques coupled with signal processing methods (Troadec 1991, Formella et al. 2007, Fisher and Hunter 2018). This usually involves reading the intensity peaks within a specific sector of the otolith, starting from the core (nucleus) down to the outer edge.

As the field of artificial intelligence (AI) has become more and more advanced, automation efforts shifted towards the use of approaches based on machine learning. Fablet and Le Josse (2005) designed one of the earliest studies utilizing machine learning algorithms to classify otolith images according to age groups. They explored the use of support vector machines (SVM) and artificial neural network (ANN) coupled with some elements of classical image processing as part of feature engineering. The work done by Bermejo et al. (2007) is another classical machine learning approach involving the use of hand-crafted morphological features combined with principal component analysis (PCA) and SVM.

Recently, with the emerging popularity of deep learning, the practice of feature engineering becomes obsolete due to the fact that this process is incorporated in the learning network itself (Bengio et al. 2013). Moen et al. (2018) became one of the earliest adopters of this technology when they used

convolutional neural network (CNN) and regression to obtain good age estimates for Greenland halibut *(Reinhardtius hippoglossoides)* otoliths.

One main issue with this existing deep learning formulation, however, is the seemingly black-box nature of the process. It is able to give age estimates, but it provides no direct information on how it derives such predictions. The follow-up studies by Ordoñez et al. (2020) and Martinsen et al. (2022) aimed to find some potential clues and explanations in the form of the heatmaps indicating individual pixel relevance. While they managed to show the focal regions considered by the algorithm, some doubts still remain as these highlighted parts are not the usual areas associated with the manual age reading process.

Another argument against the above-mentioned traditional CNN approaches is that they are known to require a large amount of training data in order to avoid overfitting. Hence, given a limited set of image data, it is possible that the implemented deep learning algorithm can only handle datasets that are very similar to those used during training. Consequently, it is very likely that the resulting deep learning model will not be robust enough to generalize and extrapolate on seemingly unfamiliar data. Recently, there have been several new studies implementing novel methods not covered in this study such as the use of transformers by Sigurðardóttir et al. (2023) and ensemble learning by Moen et al. (2023), which potentially can address the mentioned shortcomings of traditional CNN while the issues of explainability remain.

In our study, we propose to overcome these limitations by reformulating the problem and approaching it from the perspective of object detection and segmentation. That is, we directly adopt how the manual age reading process is done by explicitly performing detection and/or segmentation of annual rings which will then be automatically counted to derive the age estimates. To accomplish this, we utilize two deep learning algorithms, namely Mask R-CNN (He et al. 2017) and U-Net (Ronneberger et al. 2015), which are known for their effectiveness in detecting or segmenting, respectively, any specified region of interest on a given image.

In this proposed reformulation of the problem, we aim to reduce the level of abstraction inherent in the process and increase the explainability of the deep learning-based approach by making the procedure directly compatible with the traditional ring counting method used by humans. Also, we hypothesize that the number of required images for training will be considerably less as each image is already composed of multiple training instances in the form of labeled annual rings, which are treated as individual regions of interest. To demonstrate the plausibility of the approach, we performed several benchmarking tests that compare the overall performance of the proposed approaches against published methods based on deep learning as well as traditional signal processing. In addition, we also evaluate and compare the robustness of the methods as well as their capacity to extrapolate and adapt to new datasets.

Materials and methods

Otolith images and their corresponding age readings were provided by the Thünen Institute of Sea Fisheries and the Thünen Institute of Baltic Sea Fisheries. The image collection, as shown in Table 1, can be divided into two sets: (1) the North Sea dataset (https://doi.org/10.5281/zenodo.8341092) and (2) the Baltic Sea dataset (https://doi.org/10.5281/zenodo.8341149). The North Sea dataset consists of images from several demersal species, including North Sea cod (*Gadus morhua*), saithe (*Pollachius virens*), haddock (*Melanogrammus aeglefinus*), and whiting (*Merlangius merlangus*). To achieve higher statistical power, we only used the otolith images with ages 1–11, as this range contains enough data for both training and testing. For the Baltic Sea dataset, the otoliths are composed purely of Baltic cod (*G. morhua*). Likewise, we only used those with ages ranging from 1 to 5, as these age groups contain a sufficient number of images for the analyses. It is important to note that, in contrast to the North Sea set, the manual age readings from the Baltic Sea dataset are all validated using tetracycline markings (Krumme et al. 2020). For more details on both datasets, a table is included (Table S5) under the Supplementary Materials.

Preliminary manual checks were done on the two image datasets to ensure that no duplicates were taken and that all images were unambiguously named. Also, there were cases where some otolith images had artifacts that obscured a significant portion of the otolith. For our purposes, it is important that those are not included. Lastly, since the methods require at least one annual ring for the ground-truth preparation, images with age 0 were also excluded.

For obtaining the North Sea otolith images, it is a common practice to apply some image filters via an imaging software to make the rings more visible. Hence, for this dataset, all the images have already gone through some preprocessing for image enhancement. The Baltic images, on the other hand, were utilized in their raw states.

Data preparation and configuration of the methods

For each dataset, we used randomized subsampling to create the training set and consequently segregate the test set with the remaining out-of-sample images. As shown in Fig. 1, the raw number of images for each age group varies considerably. To avoid a prediction bias towards the age groups with more data, the subsampling was done such that there is a rebalancing of age groups after every randomized selection (i.e. given a certain quantity, excess training images on some age groups were removed while those with fewer images were refilled).

As shown in Table 2, the partitioning of the datasets was done for multiple experiments. Each age group contains the same number of training images with the exception of species-wise experiments. Lastly, apart from splitting the data into training and test sets, there is also a need to select the validation set that determines the training checkpoints (i.e. for saving the model state in each epoch whenever there is an improvement in the loss computed). Instead of further dividing the training set to create the validation set, we opted to construct it via data augmentation involving horizontal flipping of the training images.

The next step was to conduct ground-truth labeling, which is required as part of the supervised learning process. In the next four subsections, we describe separately each algorithm involved in the study to highlight their differences and some simplifications adopted for our purposes. The first two algorithms, namely classical image processing and CNN regression, represent the methods that are already existing in the literature and which serve as baselines for comparison. Then, we describe our proposed approaches based on Mask R-CNN and U-Net and elaborate the way these methods can perform

Table 1. A summary of the number of images available per species, along with the sampling area and abbreviations used in this study.

Species	Area	Number of Images
Gadus morhua (N-cod)	North Sea	194
Pollachius virens (N-saithe)	North Sea	351
Melanogrammus aeglefinus (N-haddock)	North Sea	78
Merlangius merlangus (N-whiting)	North Sea	37
Gadus morhua (B-cod)	Baltic Sea	1155

For species-wise experiments and analyses, both the N-haddock and the N-whiting were not used as they have insufficient quantities.

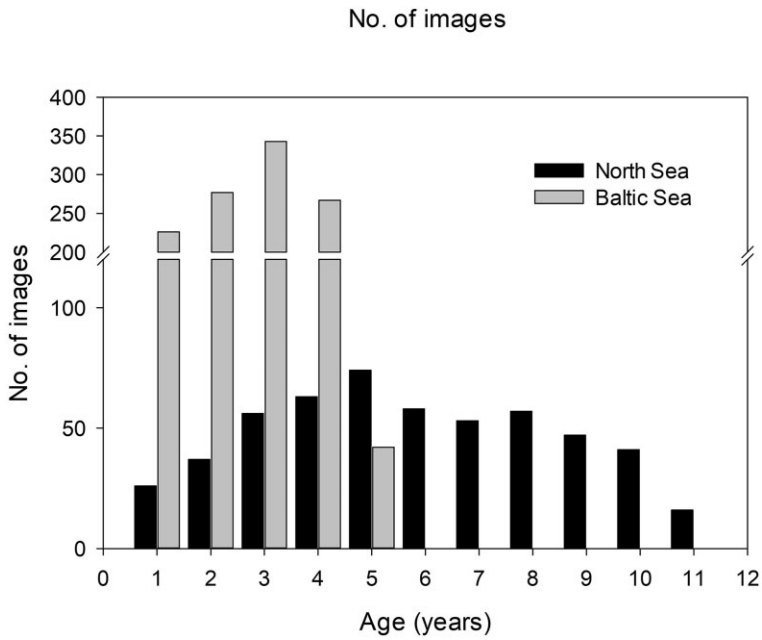

Figure 1. The number of images for each age group for both datasets. A total of 660 otolith images (ages 1–11) were included for the North Sea dataset, while there were 1155 images in the Baltic Sea dataset (ages 1–5). For a detailed tabular summary of each age group, please refer to the Supplementary Material.

Table 2. The number of images used for each data split, along with the number of runs or subsampling replicates done in each experiment.

Experiment Type	Training and Validation	Testing	Runs
Basic evaluation	132–North Sea images	528–North Sea images	20
	150–Baltic Sea images	1005–Baltic Sea images	4
Robustness test	132–North Sea images	528–North Sea images	20
	150–Baltic Sea images	1005–Baltic Sea images	4
Age extrapolation	84–North Sea images	188–North Sea images	8
	120–Baltic Sea images	42–Baltic Sea images	4
Interchanging domains	132–North Sea images	1155–Baltic Sea images	20
	150–Baltic Sea images	660–North Sea images	4
Trained with N-cod	132–N-cod images	351–N-saithe images	8
		1155–B-cod images	8
Trained with N-saithe	132–N-saithe images	194–N-cod images	8
		1155–B-cod images	8

The validation data is derived entirely by data augmentation of training images via horizontal flipping operation; it hence has the same quantity as the training set.

age estimation totally compatible with traditional ring counting methods. To facilitate the understanding of the entire process, our source code (written in Python 3.8 (Van Rossum and Drake 2009) with machine learning libraries such as Keras 2.2.4 (Chollet et al. 2015) and Tensorflow 1.15 (Abadi et al. 2015)) is available on Github (https://github.com/arjaycc/ai_otolith/tree/v1.2). Also, a schematic diagram outlining the main steps for the proposed deep learning approaches is given as a Supplementary Material (Fig. S12).

Classical image processing

For the image processing approach, we chose to explore mainly the methods that use intensity peak counting, as this approach is quite popular and straightforward to use, as reviewed by Fisher and Hunter (2018). Simplifying the ideas from the literature (Troadec 1991, Formella et al. 2007), the method we finally implemented was to simply create a polar transformation of the sector slices from otolith images and convert them into square tiles using the relative distances of

the pixels starting from the otolith nucleus or core down to the outer edge. A schematic diagram of the process is given in Fig. S15in the Supplementary Materials.

As a preliminary step, we needed to first identify the outer otolith contour and the nucleus from the images. A simple application of the watershed algorithm [from the Python Skimage Library (Van der Walt et al. 2014)] isolates most otoliths from their corresponding background with great accuracy, from which the outer contour can be obtained. There are a few cases that appear to generate erratic contours, especially if the outer otolith edges are not clearly distinguishable. For our purposes, we simply identified and manually corrected these erratic contours by using a standard image annotation tool. We opted to use the Visual Geometry Group (VGG) Image Annotation tool abbreviated as VIA (Dutta and Zisserman 2019), due to its simplicity and extensibility. In fact, we managed to incorporate our own code into this tool, where we created a brush feature to facilitate the annotation, as it is also needed for the ground-truth preparation of the other methods.

For identifying the nucleus of the otoliths, several classical image processing techniques are also widely popular (Fablet and Cao 2006, Harbitz 2009). We chose a simple heuristic based on ellipse approximation (Harbitz 2009) to locate the approximate nucleus position, which worked quite well for the Baltic Sea dataset. However, for North Sea images, some nucleus coordinates were missed, so we had to do manual adjustments using the same annotation tool so as not to introduce another source of error and to focus only on the steps involving annual rings.

Overall, the entire process relies on the assumption that there is a proportionality among the growth of the rings on a certain local portion of the otolith (Fablet and Le Josse 2005). Hence, it is expected that when the otolith sectors are sliced and divided into small enough pieces, the transformed rings will be approximately aligned (Fig. 2a). With these transformed images, it is straightforward to generate a good intensity signal plot by taking either the mean or median of pixel rows from top to bottom across multiple slices along the major axes (Fig. 2b). To derive the age reading, we performed a peak counting procedure using a peak detection algorithm based on a standard implementation available from the literature (Billauer 2009).

CNN regression

CNNs are one of the most widely used algorithms to deal with image datasets (Krizhevsky et al. 2012). The core idea is roughly inspired by the biological neural network, where the concept of neurons is represented using mathematical interconnected nodes (O'Shea and Nash 2015). The information propagation is made through a process of weight updates along these interconnected nodes using intricate mathematical operations with the goal of making the predictions be as close as possible to the actual or expected value through the evaluation of one or more loss functions during each training epoch. These nodes are typically grouped into layers and each node can have multiple connections into other nodes located at the next layer. What primarily differentiates CNNs from traditional ANNs is the number of layers; for the former, it is several orders of magnitude higher (i.e. the layers go deeper) than for the latter.

The most basic use of a CNN is in a supervised manner, which could be formulated as either classification or regression (Martinsen et al. 2022; Moen et al. 2018, Ordoñez et al. 2020, Politikos et al. 2021). That is, a discrete or continuous value will be returned as a prediction, which directly corresponds to the probable category or measurement that it learned from the labeled training data. In the case of regression, a basic loss function for the CNN is usually in the form of mean squared error (MSE) (Martinsen et al. 2022; Moen et al. 2018), which is given in the following equation:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

For the CNN method used by Moen et al. (2018), they chose to use regression, where the age estimates are turned into a continuous value. Also, they used another useful concept of CNN known as transfer learning, where a pre-trained model, primarily InceptionV3 (Szegedy et al. 2015), was reused by preloading its weight into the network prior to training.

To use this approach, mainly as a benchmark reference, we obtained the exact implementation from Moen et al. (2018), available at https://doi.org/10.21335/NMDC-1949633559. It only involves a simple data loading step where a list of image paths and their corresponding age labels are placed in a comma-separated file. This file will then be taken by the algorithm to start the supervised training. One minor issue, however, is their use of otolith pairs (left and right otoliths), which is not applicable in our study. Therefore, as a simple workaround, we flipped each otolith image horizontally to somehow have a pseudo-pairing and make the implementation compatible.

Mask R-CNN

Likewise, as implied by its name, Mask R-CNN is also a deep learning algorithm based on CNNs (He et al. 2017). The main output, however, is primarily in the form of detection masks and bounding box coordinates of the object of interest as found within the image. That is, detection masks are pixel markings that indicate the spots occupied by the object of interest, while the bounding box consists of numeric coordinates within the image that contains this object of interest. This feature of the Mask R-CNN algorithm allows it to perform both object detection and instance segmentation simultaneously.

For this algorithm, there is an implementation from Matterport (2017) containing the entire learning workflow, starting from the data loading step up to the training as well as the testing. To utilize the code, we first need to provide its needed inputs, namely the image and its ground-truth annotations. As discussed above, we selected the VIA tool (Dutta and Zisserman 2019) for annotating the images due to its simplicity. We marked the parts of the images along the left and right major axes that represent a portion of the winter annuli to be treated as the objects of interest for detection. Figure 3 shows an example of the annotation using the VIA tool.

As mentioned above, there can be different ways to implement loss functions for each algorithm. For Mask R-CNN, instead of the basic MSE, it needs to have multiple loss functions in order to check how far the predicted masks are from the actual regions while also computing the errors for the predicted bounding boxes (He et al. 2017). In the study by Zimmermann and Siems (2019), they further added another loss function related to the edges of the contours generated from the predicted masks. We used this version since it was demonstrated to learn faster and more efficiently (Zimmermann and Siems

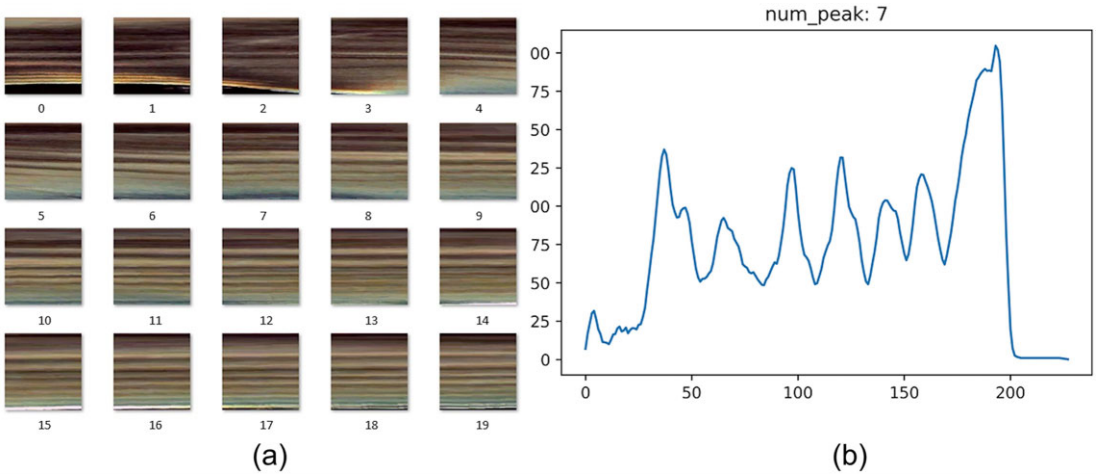

Figure 2. (a) A set of image strips that were transformed from otolith sector slices along a reading axis. (b) The resulting intensity plot when the row-wise average was taken for a single strip with the resulting peak count at the top.

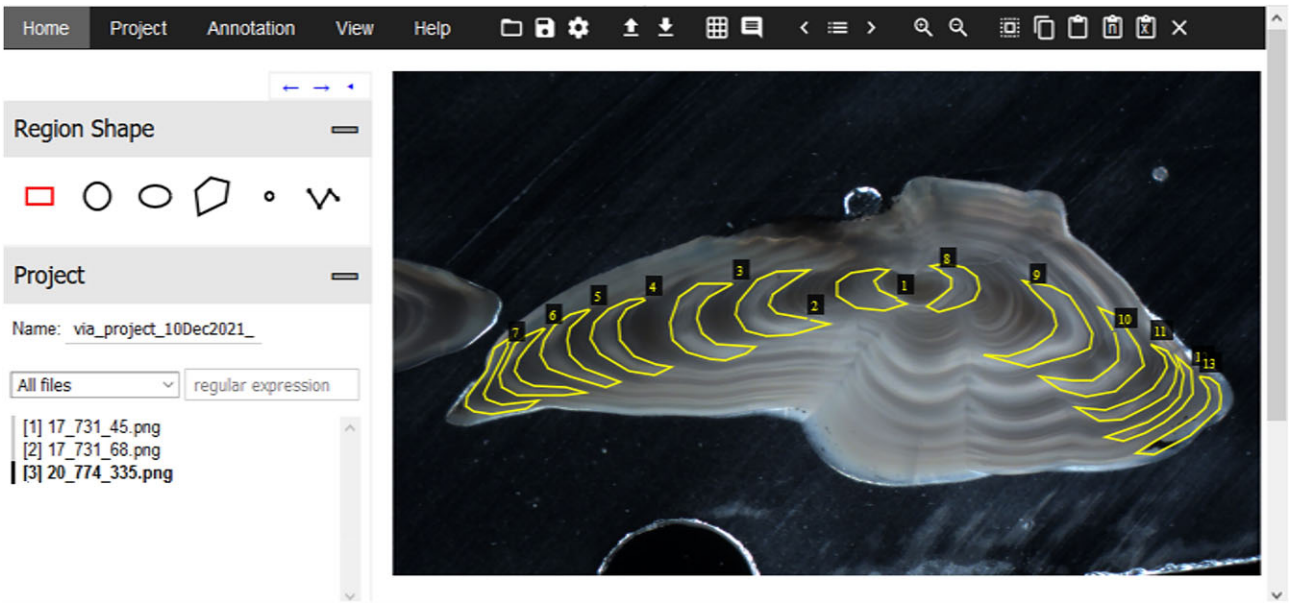

Figure 3. An example of ground-truth preparation made using the VIA annotation tool showing the annotations (yellow) that mark the regions of interest within an otolith image.

2019). Also, for this implementation, transfer learning was involved, where an existing model (Matterport 2017) trained from the COCO dataset (Lin et al. 2014) was preloaded instead of training from scratch.

The output of Mask R-CNN still needs to undergo a post-processing step in order to derive the age estimates. A schematic diagram was included in the Supplementary Materials (Fig. S16) that summarizes the process. The core idea is to scan the masks and to find their alignment towards the center, which indicates that they belong in the same reading axis. The process starts by locating the nucleus and measuring the distances of the masks to this reference point. Then, each mask is visited from the nearest to the farthest to label their positions. To perform labeling, the angle (in radians with respect to the nucleus) of a mask is measured through its endpoints. Two masks are aligned if their angles overlap. To label a mask, increment by 1 the label of the most recently visited mask that aligns to it. If there is none, then label it as 1. Once all the masks are visited, sort the labels, then find the highest value, which will indicate the highest ring count (corresponding to the age reading).

U-Net

U-Net also makes use of the CNN architecture (Ronneberger et al. 2015), similar to the two previously presented deep learning algorithms. The main difference, however, is that for U-Net, the final output is composed only of a segmentation mask for the entire image, corresponding to the pixels detected representing the object of interest. Because of this, U-Net is usually utilized for problems involving the semantic segmentation of images.

To train the algorithm, ground-truth masks are likewise needed to mark the regions to be segmented by the U-Net. In this study, two ways of ground-truth labeling were followed. One method involved masking the entire concentric annuli and the other involved masking only a certain portion

of the annuli along the reading axes. For the former, new sets of ground truth masks have to be created using the VIA annotation tool. For the latter, we reused the same reading axes annotations made previously for Mask R-CNN.

In contrast to Mask R-CNN, only the segmentation masks are returned by U-Net and no bounding boxes are generated. Hence, there is usually only one loss function involved, which determines whether each pixel of the image was properly marked either 1 or 0, depending on whether they are part of the object of interest or not, respectively. The basic loss function can be the MSE of these per-pixel differences, but it can be modified as needed. In fact, for this study, we used the extension proposed by Ronneberger et al. (2015), where more weights are given on the pixels in between the objects of interest. That is, the algorithm has to be more careful when marking those pixels between each annulus because the errors from these portions weigh more than the rest. Otherwise, without weighted loss, the U-Net has the tendency to return overlapping contours, especially when the rings are near each other, which is particularly happening at the edges.

Similar to Mask R-CNN, the segmentations cannot be readily isolated from the rest of the pixels. Hence, it is therefore also necessary to perform a post-processing step in order to remove noise and easily count the proper segmentations where age estimates are derived. This process is summarized in Fig. S16 of the Supplementary Materials. As there are no detection scores like the ones from Mask R-CNN, we applied a simple noise filter based on the size and relative position of the segmentation. That is, if a segmentation is found, it is first checked whether it is just a random noise before including it in the ring count. This is similar to the criteria also applied in peak detection methods, where certain peaks are eliminated according to their relative sizes and positions.

In this study, we explored different configurations for this algorithm in order to identify the best-performing variant. First, two different ground-truth methods were tested: one annotation set marks only the portion along the major axes, while the other annotation covers as many annuli as visible in the image. Secondly, we also compared the performance of implementing U-Net with transfer learning using pre-trained VGG weights (Simonyan and Zisserman 2015), similar to the implementation of Abdellatif (2021), against the default implementation, which is trained from scratch. At this point, it is worth mentioning that for all the pre-trained models used in each deep learning method explored in this study, the training set from which they were originally trained on are all composed of images from common objects and not specifically for otolith.

Benchmarking

There are three basic benchmark tests that we conducted in order to thoroughly assess and compare the overall performance of the algorithms, which we measured in terms of percentage agreement. For the first test, we performed the usual training, validation, and testing using images from the same data source. This test also involved identifying initially the best hyperparameters and configurations of each algorithm that would be used for subsequent experiments. For the CNN-regression method, we used the default or suggested hyperparameters taken from the study of Moen et al. (2018). For the other algorithms, we implemented a simplified grid search on the different configurations and hyperparameters and evaluated their performance on a subset of the test data.

For the second test, we performed some variations of the first test to evaluate two criteria: (1) the robustness of the algorithms when slight changes/perturbations on the images are introduced, and (2) the ability of the algorithms to extrapolate (higher) age groups when they are explicitly removed from the training data. The former involved simple background removal with increased brightness on the test images to see whether the algorithms have taken cues on unreliable features such as background artifacts or even the differences in lighting. The latter involved the complete removal of any training data from higher age groups (ages 8–11 for the North Sea dataset, age 5 for the Baltic Sea dataset) to see if the algorithms could extrapolate these higher age ranges without encountering them during training.

Lastly, for the third batch of tests, we checked for inter-dataset and inter-species performance to assess how adaptable the models are when analyzing new sets of data from a completely unfamiliar species or domain. For the basic case, we interchanged the test sets for North Sea and Baltic Sea otoliths and assessed the new performance (i.e. the models trained from North Sea images were tested against Baltic Sea test images and vice versa). For the other case, we segregated the images further into different species to see whether training them on a specific species makes the algorithms completely unable to generalize on the other species. Conversely, we also aimed at finding out whether training the algorithms on a given species allows them to handle the same species from a completely different source. For simplicity, in this experiment, we use the term inter-species loosely, despite also treating the North Sea cod and Baltic Sea cod as separate groups.

Coefficient of variation analysis

In the context of age reading evaluation, apart from percentage agreement, another important metric is the so-called coefficient of variation (CV), which is especially useful during age reading workshops where readers from various institutions gather to cross-check the possible differences in the way they perform age readings. This value can be computed using the following formula (Campana 2001):

$$CV = \frac{\sigma}{\mu} \cdot 100,$$

where σ = standard deviation and μ = mean of age estimates from the readers.

For reference, we used two separate ICES workshops, one for North Sea cod (2008) and another for Baltic Sea cod (2020), where participating readers performed age estimation on cod images using their own methodologies. It was reported that for both the North and Baltic Sea workshops, the readers had a significant disagreement indicated by the computed CV of about 40% (39.8% to be precise) and 15%, respectively. Optionally, for the North Sea workshop, we may exclude the values contributed by broken otoliths and refer only to the result for sectioned otoliths, which is around 22.5%. Hence, for this study, similar to the formula used by Moen et al. (2018), we also computed the CV by treating the automated and manual readings as individual readers and assess whether the age estimate variations fall within the same range attained by human readers.

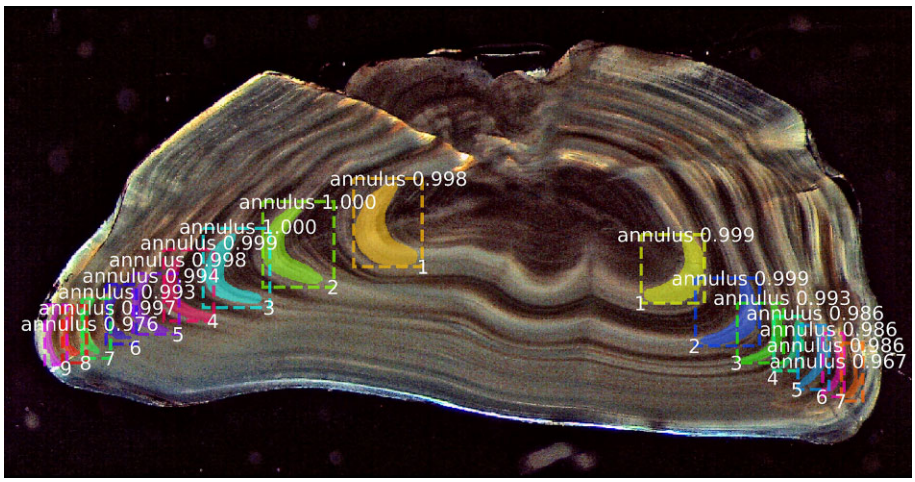

Figure 4. Mask R-CNN object detections with the corresponding bounding boxes and scores. Higher scores indicate higher model confidence, which can be used to filter out those predictions that do not surpass a certain detection threshold (i.e. a hyperparameter that can be adjusted as needed).

Statistical analysis

To check for the statistical significance of the comparisons, we used the standard pairwise *t*-test available in the R programming language (R Core Team 2020) along with the correction proposed by Nadeau and Bengio (2003), which is implemented in the correctR package (Henderson 2023). We carefully considered the fact that some assumptions of the standard *t*-test are violated by the data partitioning used to create the training and test splits. As mentioned earlier, we employed a small variation of the randomized subsampling for creating the training and test sets, which means that the images used for each run are not completely independent (i.e. the training and test sets of one run could have images that were also included in the other runs). This leads to a high probability of type I error causing a problematic rejection of the null hypothesis in pairwise comparison of algorithms (Dietterich 1998). Fortunately, the ground-breaking study made by Nadeau and Bengio (2003) suggests that a simple correction of the standard *t*-test can overcome this limitation. Therefore, for the main test involving general performance comparisons, this corrected resampled *t*-test is used as it satisfies the conditions needed for the statistical analysis. For the other test cases which deviate greatly from standard randomized subsampling (e.g. age-wise and species-wise test), we used the standard *t*-test while taking into account the potential pitfalls mentioned.

Results

One straightforward advantage of the CNN-regression algorithm used by Moen et al. (2018) is that the age readings are readily available and directly outputted in the model predictions. For all the other methods, however, an intermediate output has to be generated first before the actual age reading can be derived.

For the classical image processing approach, the intermediate results are in the form of signals that indicate the image intensity values from the nucleus to the outer edge of the otoliths as shown in Fig. 2b.

For the Mask R-CNN, the final detections need to be post-processed first as described in the "Materials and methods" section in order to directly appear on the image as shown in Fig. 4. Apart from the colored masks, it can be seen that there are bounding boxes that are also depicted containing the prediction scores. These values range from 0.0 to 1.0 and directly correlate with the model's confidence on the predictions.

For the U-Net algorithm, the intermediate result also needs to undergo post-processing before the age estimates can be derived. Figure 5 shows an example of a raw mask output of the U-Net as well as the resulting image masks after the post-processing procedure similar to the one performed for the Mask R-CNN output.

After the post-processing stage for each algorithm, the derived age estimates are then plotted against the manual age readings, as shown in Fig. 6. It can be seen that there is a diagonal trend that becomes apparent with these plots, indicating the relative agreement, between the automated and the manual readings. The plot also shows how far the under- and over-estimates are from the diagonal, indicating the biases of each method. For illustration, only the test results of a single run with North Sea images are shown in the figure. For the plots of all the runs, including those of the Baltic Sea images, refer to the Supplementary Material.

Figure 7 provides a clearer comparison of the performance of the different algorithms tested. The resulting trend is different for the North Sea dataset and the Baltic Sea dataset. The CNN regression has a clear edge with 55% and 87% mean accuracy for North Sea and Baltic Sea images, respectively. The Mask R-CNN has a slightly poorer performance on North Sea images (46%) but it has a decent mean accuracy on Baltic Sea images (72%). On the other hand, the U-Net algorithm manages to be competitive with 54% mean accuracy on the North Sea dataset and 72% mean accuracy for the Baltic Sea dataset. Lastly, the traditional automation method using classical image processing attains the poorest performance, showing only 26% and 54% mean accuracy for the North Sea and Baltic Sea datasets, respectively. Hence, this approach was no longer used for further analysis to focus more on the deep learning algorithms.

Using the corrected resampled *t*-test, the null hypothesis that CNN-regression results do not differ from the results of both the proposed methods has failed to be rejected in the North Sea dataset (P-values > 0.05), while it was rejected for the Baltic Sea dataset (P-values < 0.05). This indicates that the

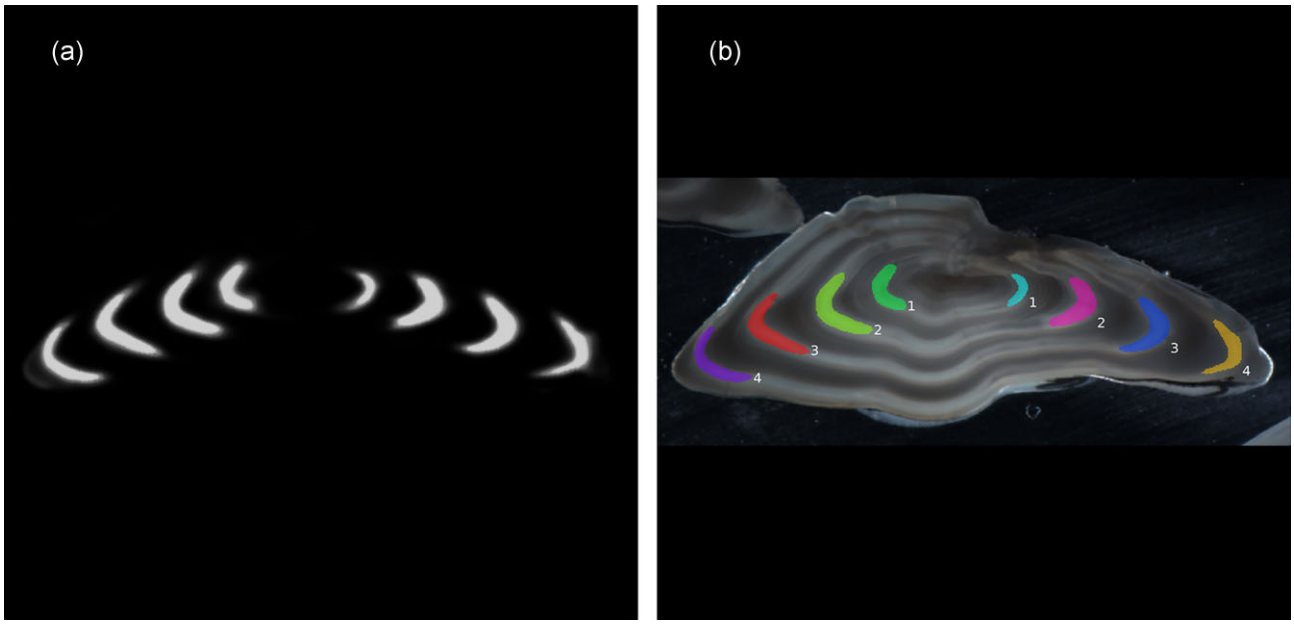

Figure 5. The raw U-Net output alongside a sample end result after the post-processing step. The direct output of a U-Net model is a mask indicating the regions it segmented (a) that can be post-processed to generate the ring count (b).

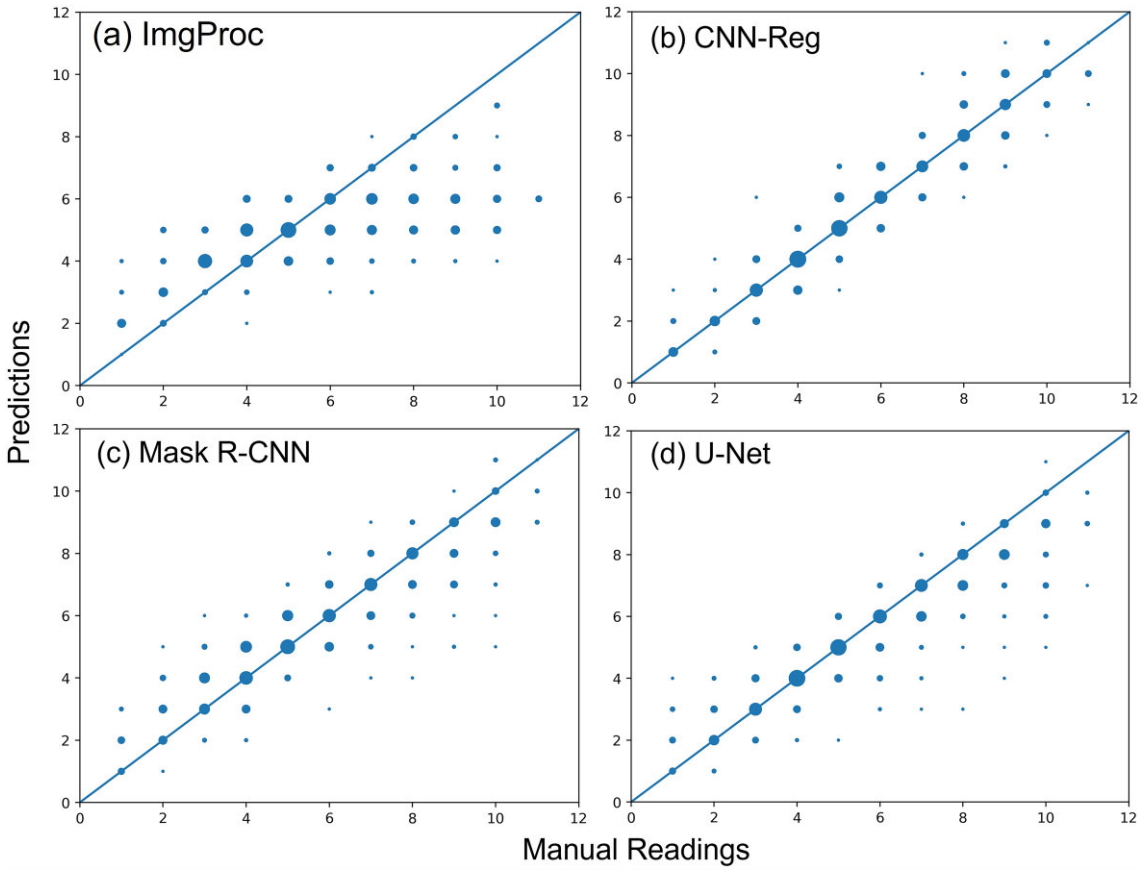

Figure 6. The plots of automated age estimates against the manual age readings on a test set involving North Sea images using the various approaches, namely (a) image processing, (b) CNN regression (rounded off), (c) Mask R-CNN, and (d) U-Net.

proposed methods have a similar performance to the CNN regression on the North Sea images but fail to attain the same competence on the Baltic Sea images, where the CNN regression shows its clear advantage.

To assess if an automated method is good enough to be treated like an individual human reader, we also computed the CV for each method as shown in Table 3. With a reference value of 40% and 15% taken from the North Sea

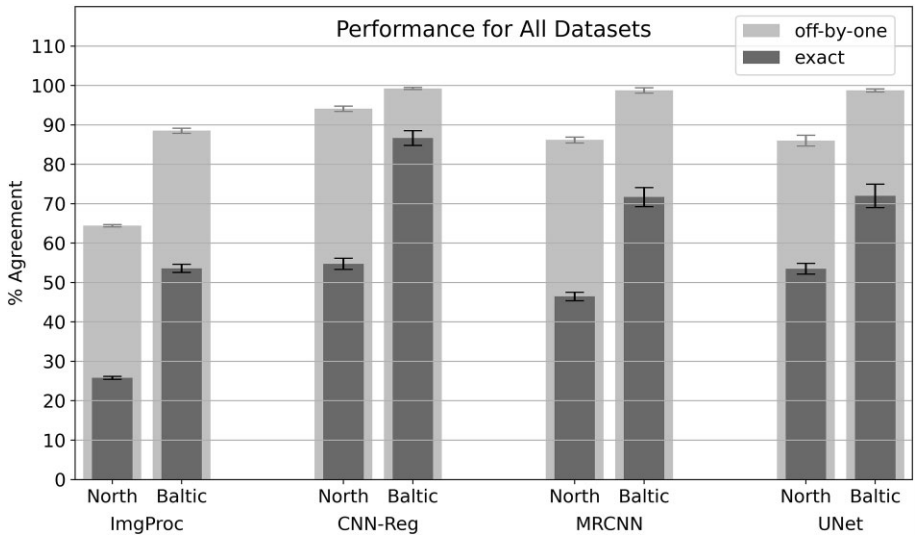

Figure 7. Overall performance of the different algorithms on the North Sea and Baltic Sea datasets across multiple runs with randomly subsampled test sets ($n = 20$ for the North Sea dataset, $n = 4$ for the Baltic Sea dataset). Applying the corrected resampled *t*-test to compare each proposed deep-learning method (M-RCNN and U-Net) to the published CNN-regression method yields corresponding *P*-values = 0.14 and 0.43 (> 0.05) for North Sea images and *P*-values = 0.003 and 0.048 (< 0.05) for Baltic Sea images.

Table 3. The coefficient of variation (CV) of the different methods against the manual readings.

Method	North Sea dataset	Baltic Sea dataset
ImgProc	19.1%	16.4%
CNN-Reg	7.4%	3.8%
M-RCNN	10.9%	10.1%
U-Net	10.5%	9.6%

The reference value is 40% for the North Sea dataset and 15% for the Baltic Sea dataset, which correspond to the CVs from a group of readers during two ICES workshops on cod otoliths.

and Baltic Sea workshops, respectively, it can be seen that the computed CVs for the deep learning methods fall significantly below these thresholds, indicating that they are indeed already at the level of human readers. It is important to note that for the North Sea workshop, we may only consider the results for sectioned otoliths and ignore the values for broken otoliths, which is not relevant in this study. Hence, even if the reference value is adjusted to 22.5%, the same conclusion is still valid. That is, the CV results from this study still fall below the workshop reference values. This means that theoretically, if the AI-based methods are included in a workshop with human readers, the readings they provide will deviate within the same range as the ones from the human readers.

The next set of experiments evaluates the robustness of the different methods when the test images are subjected to slight variations (i.e. involving background removal and increased brightness). Figure 8 reveals one surprising disadvantage of the published CNN-regression method compared to the proposed methods. Just with the mentioned image perturbations, a very drastic change in performance is seen for the CNN regression in both the North and Baltic datasets. Only a slight degradation of performance is observed for Mask R-CNN and U-Net.

Another interesting experimental setup was designed to measure the ability of the methods to extrapolate on the data they had not encountered before. In this experiment, we removed the training images with high age values and limited the range to ages 1–7 for the North Sea dataset and ages 1–4 for the Baltic Sea dataset. Then, we tested the resulting models on a test set containing only images with age values greater than those used during training. Figure 9 summarizes the result and demonstrates the extrapolation abilities of the different methods.

It can be immediately seen that the published CNN regression fails almost completely in getting any correct estimate for higher age groups that were not included during training. In contrast, both the proposed algorithms manage to attain a decent accuracy level, showing their ability to extrapolate on unknown data.

For the last test, we further highlighted the capacity of each algorithm to handle datasets that were not introduced during training. For the first case, we interchanged the test images of both datasets and re-tested the previously trained models without re-training on the new set. That is, the existing models trained from the North Sea dataset were tested on the Baltic Sea test images, and vice versa. Figure 11 demonstrates yet another advantage of our proposed algorithms compared to the published CNN-regression method.

Overall, it can be seen that the CNN-regression algorithm attains the worst performance when given a new and unfamiliar data source or domain. This means that it learned features too specific on the dataset it was trained on resulting to its failure to generalize on the other dataset with seemingly new otolith characteristics, different microscopy lighting, and image capture techniques. In fact, this concept , referred to as domain adaptation, has also been explored in the study by Ordoñez et al. (2022), where they also evaluated this capacity on a similar standard CNN implementation but with classification instead of regression. They used images of the same species (Greenland halibut) from two different sources: one dataset came from the Norwegian laboratory, while the other dataset was taken from their counterpart in Iceland. Similar to what we have observed, they also reported that this standard CNN formulation performed poorly when tested across

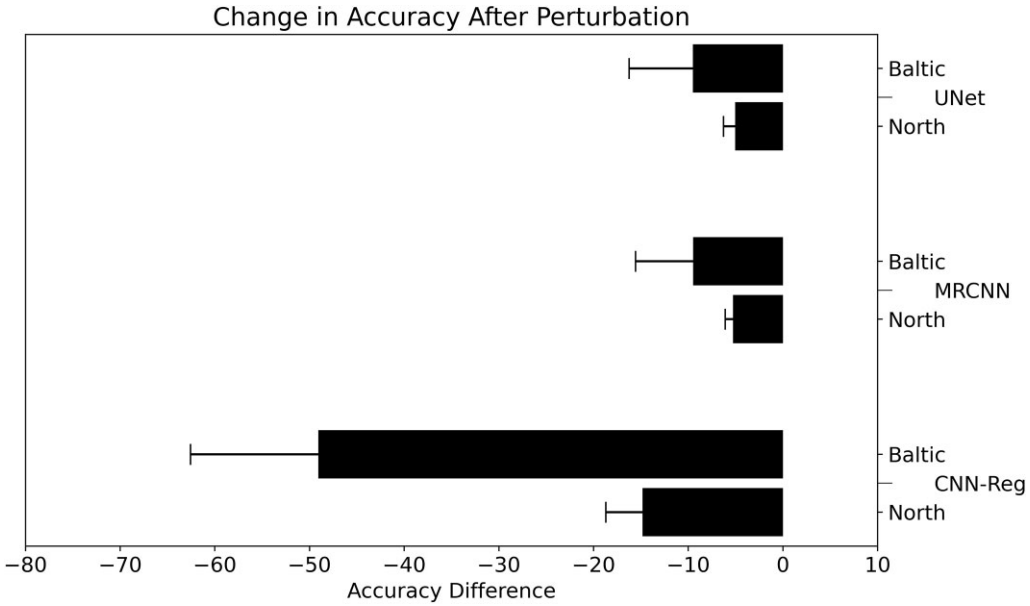

Figure 8. Degradation of the predictive performance of each algorithm when the background of the otoliths on the test images is removed while subsequently increasing the image brightness. Comparing the changes in accuracy of the proposed methods against that of the CNN-regression yields *P*-values < 0.05 using a standard *t*-test ($n = 20$ for the North Sea, $n = 4$ for the Baltic Sea).

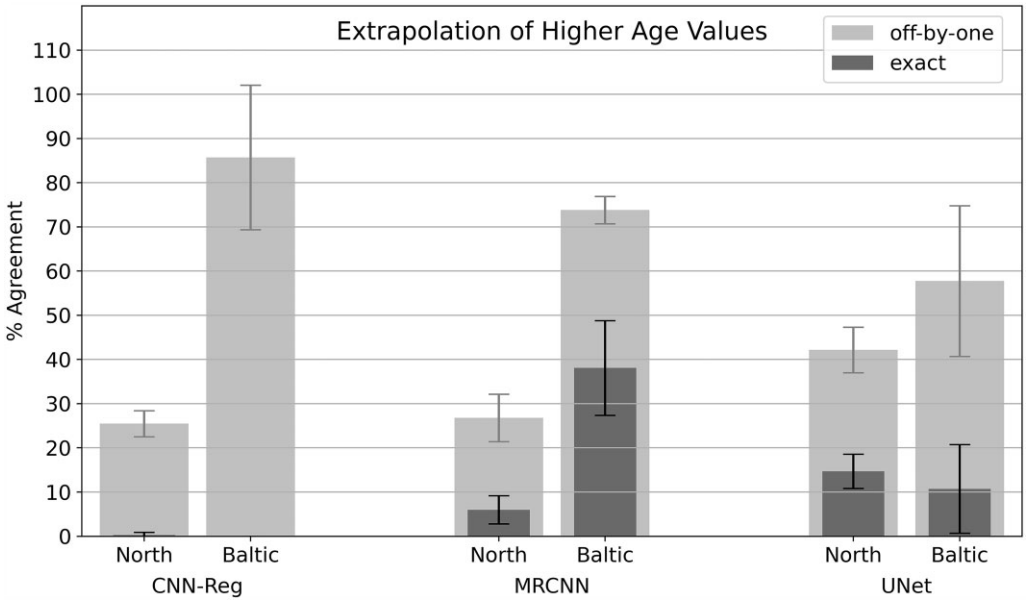

Figure 9. Performance of each deep learning algorithm on higher age groups that were excluded during training. For the North Sea runs ($n = 8$), images with ages 8–11 were used for testing as they were excluded from training. For the Baltic Sea runs ($n = 4$), only age 5 images were left out during training and were consequently used for testing. The standard *t*-test gives *P*-values < 0.05 for the pair-wise comparison against CNN regression.

the two different data sources. Hence, they proposed certain modifications to the default implementation, but this is beyond the scope of our study.

To elaborate on this observation further, we conducted another test focusing mainly on inter-species performance. For this setup, we explicitly trained the algorithms using only one specific species and performed tests on the other species. Figure 11a shows the comparison of test performance across species when the training involves only North Sea cod images, while Fig. 11b shows the results if only North Sea saithe images were included.

There are some interesting observations worth emphasizing for this batch of results. First, it can be immediately seen from both plots that the overall inter-species accuracy of the proposed methods surpasses that of the previously published CNN-regression method, indicating that the proposed methods have more generalization capacity. Specifically, the performance discrepancy is quite large when it comes to the Baltic test images. This is somehow surprising when compared to the result from the previous experiment. It seems that purely using North Sea cod images for training makes the performance of the published CNN-regression method to become

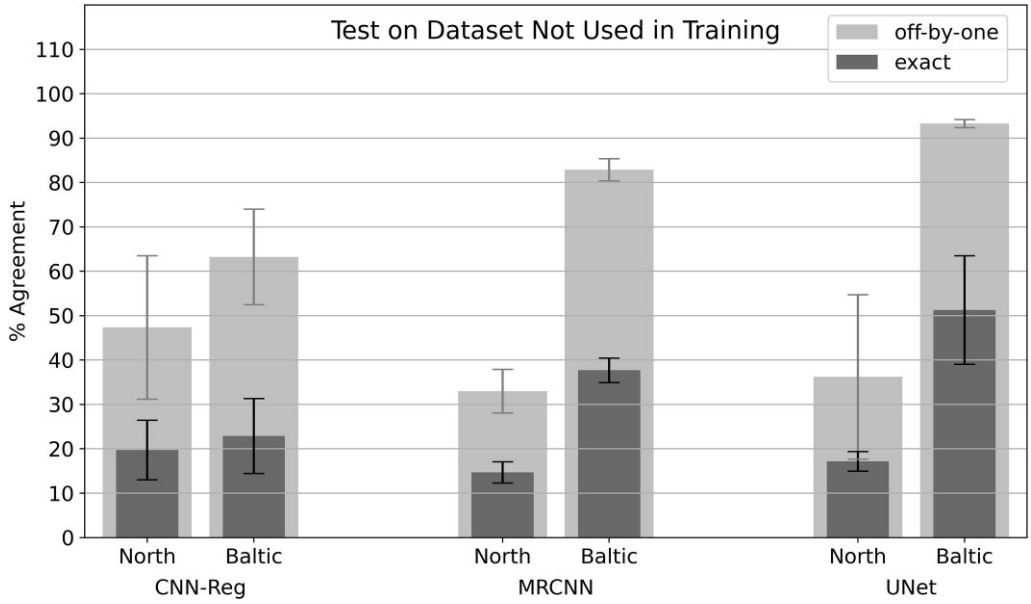

Figure 10. Performance of the deep learning models trained on one dataset and tested against the other dataset and vice versa. For the Baltic Sea test case ($n = 20$), the standard t-test show significant difference ($P < 0.05$) when comparing the CNN regression against the proposed methods.

even worse compared to using a mixed set (Fig. 9) or even pure North Sea saithe images (Fig. 11b). This result is directly in contrast to the results from the two proposed methods, where the accuracy values for predicting a new set of images coming from a different source (e.g. Baltic dataset of purely cod) become higher when the training set involves the same species (i.e. North Sea cod in Fig. 11a) compared to a completely different species (i.e. North Sea saithe in Fig. 11b). This implies that there could be species-specific patterns utilized by the proposed algorithms to help in the prediction of a new set of the same species.

In summary, from Figs 9, 10, and 11, it can be concluded that the CNN-regression method exhibited the least adaptability when it was subjected to a completely unfamiliar dataset. This means that to use this algorithm for each new species or even just a new age group, a new batch of training has to be performed to update the model, or, in the worst case, a complete retraining has to be conducted to create a totally different model. In contrast, for the two new algorithms proposed, the previous knowledge they had on one species can potentially still be usable for another species.

Discussion

Various studies have already shown that the standard CNN classification or regression performs satisfactorily when it comes to age estimation of various fish species (Moen et al. 2018, Politikos et al. 2021, Martinsen et al. 2022). Apart from the predictive power, another big advantage of their approach is the training simplicity, where minimal ground truth preparation is needed. However, to be widely accepted, this formulation has one big issue, and that is its black-box nature. The follow-up study done by Ordoñez et al. (2020) tried to find a way to explain the decisions for this type of CNN but it still leads to more questions and counter-intuitive observations.

In the work presented here, we have shown that the use of object detection and segmentation algorithms can be a good alternative formulation when it comes to automating the fish age reading process. In addition to having a comparable performance on multiple test sets, we demonstrated that it also has several advantages compared to multiple methods that can be found in the literature. In particular, we showed that the resulting models are more robust even when some perturbations are introduced into the images. Also, we demonstrated its ability to extrapolate and generalize on datasets that were not introduced during the training phase, especially those coming from a completely different source. Lastly, and maybe most importantly, this new way of applying deep learning on automated age reading makes the overall process more explainable due to its direct compatibility with traditional manual methods.

One major drawback is the seemingly tedious process of doing data preparations, especially the ground-truth labeling. While this may be true, it is important to note that this will only be the case if we need to train a new model with each new dataset that we obtain. However, as demonstrated by the results, there is a potential for the object detection and segmentation models to be reusable with a completely new dataset. This means that the ground-truth preparation will eventually become less and less required as retraining becomes unnecessary in some instances. In contrast, the standard CNN regression formulation will always need to be trained with each new dataset due to its lack of adaptability.

It is important to note, however, that all these observations involving the CNN regression formulation are only tested using the implementation from the study conducted by Moen et al. (2018). It is possible that with newer designs and architecture, these limitations may no longer be true. Also, there are already novel approaches that exist in the literature that seem promising when it comes to handling the known limitations of older deep learning designs, such as the use of transformers (Sigurðardóttir et al. 2023) and ensemble learning (Moen et al. 2023). It will indeed be interesting to conduct further benchmarking with these new approaches to see if the advantages of our proposed methods remain valid. Also, it is worth mentioning that the statistical tests performed in this study,

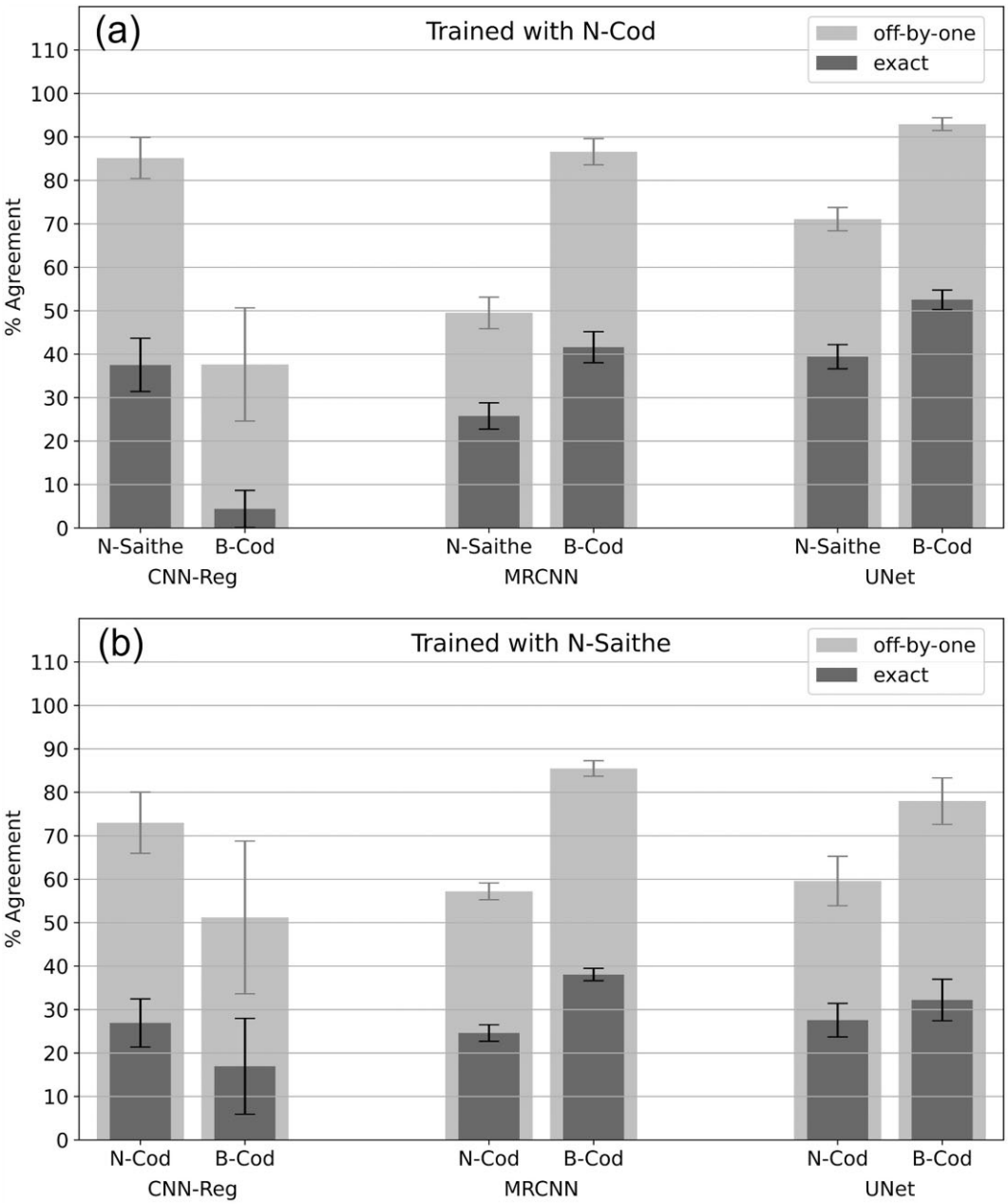

Figure 11. Performance of the methods across species (and stock) when the training involves (a) only North Sea cod images. (b) only North Sea saithe images. The standard *t*-test ($n = 8$) shows high significance ($P < 0.05$) on the Baltic cod test case for both Mask R-CNN and U-Net after pairwise comparison against the previously published method.

namely the standard *t*-test and the corrected resampled *t*-test, have limitations with respect to reducing type I and type II statistical errors (Nadeau and Bengio 2003, Bouckaert and Frank 2004), so more repetitions are needed to make stronger claims. It is hence an option to explore other statistical methods apart from a *t*-test, which will ensure that both the type I and type II errors are minimized during benchmarking.

Lastly, one important concept of CNN that is widely used in this study is the concept of transfer learning. For all the deep learning approaches we tested, we took advantage of this facility and preloaded some pre-trained models. Therefore, there is an apparent future direction where the process of reusing a newly trained model can be improved further and training can be done using a base pre-trained otolith model (instead of VGG16 or InceptionV3). Also, for U-Net and Mask R-CNN, this base model can possibly aid on generating new ground-truth labels for future datasets and then enable a self-sustaining loop where each updated model will be reused to generate annotations for newer datasets and so on. In this way, the creation of annotations will be AI-assisted and not entirely done from scratch, needing only a simple manual correction if necessary.

Conclusion and future outlook

With the growing size of the otolith image datasets that are being collected and processed by various institutions, it is becoming apparent that the advances in the fields of big data an-

alytics, computer vision, and machine learning can be of great use. This study is another step towards scalable otolith analysis, and it successfully demonstrated how one can utilize the well-known techniques in object detection and segmentation to automatically perform age reading on otolith images.

As the age estimates of AI-based methods match closer and closer to those from manual age readings, it becomes clearer that the predictive performance is not the only criterion towards their general acceptance. Features such as robustness, adaptability, and, in particular, explainability are also important considerations, which were all exhibited by the proposed approaches in this study.

With an automated system for age estimation, the process of analyzing a large number of images can be highly efficient, scalable, and less susceptible to logistic and subjective limitations. Using the proposed algorithms, we aim to create a framework or a system (i.e. a web application) that can be used as a platform for high-speed processing of large datasets. As a general toolkit for otolith image analysis, it can be made to provide not only age information but also other relevant measurements such as otolith radius and annulus distances, which are useful parameters for certain biological and ecological models. Lastly, we also hope that this future framework can be an avenue for a more collaborative effort within the community where models, images, and even annotation data can be shared efficiently and even allow continuous enhancements of existing models and techniques.

Supplementary material

The following Supplementary material is available at *ICES Journal of Marine Science* online.

Funding

This study was supported by the Thünen Institute.

Acknowledgements

We thank the members of the Otolith Age Reading Group at Thünen Institute of Sea Fisheries for the detailed discussions and walkthrough of the process of manual age reading. We thank Friederike Beußel and Hendrik Brückner for the collection of North Sea otolith images. Likewise, we thank Dr Uwe Krumme from the Thünen Institute of Baltic Sea Fisheries for providing the Baltic Sea dataset. Lastly, we thank Marianne Camoying for the helpful insights on manuscript writing.

Author contributions

All authors were involved in the conceptualization of the experiments as well as the formal analysis of the results and conclusions of the study. A.C. wrote the original draft of the manuscript and all authors contributed to the revisions and the overall review and editing. .

Conflict of interest: The authors declare no conflict of interest.

Data availability

The source code of the study is publicly available at https://github.com/arjaycc/ai_otolith/tree/v1.2 and can also be downloaded from https://doi.org/10.5281/zenodo.8341297. The datasets used, however, need to be downloaded separately at the following locations: https://doi.org/10.5281/zenodo.8341092 for the North Sea dataset and https://doi.org/10.5281/zenodo.8341149 for the Baltic Sea dataset. Lastly, a subset of the trained models from the study can be downloaded at https://doi.org/10.5281/zenodo.10000645.

References

Abadi M, Agarwal A, Barham P *et al.* TensorFlow: large-scale machine learning on heterogeneous systems. *arXiv:1603.04467*, 2015.

Abdellatif AH. https://www.kaggle.com/code/aithammadiabdellatif/vgg16-u-net. 2021. (March 2020, date last accessed).

Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 2013;**35**:1798–828. https://doi.org/10.1109/TPAMI.2013.50.

Bermejo S, Monegal B, Cabestany J. Fish age categorization from otolith images using multi-class support vector machines. *Fish Res* 2007;**84**:247–53. https://doi.org/10.1016/j.fishres.2006.11.021.

Billauer E. peakdet: peak detection using MATLAB (non-derivative local extremum, maximum, minimum), 2009. http://billauer.co.il/peakdet.html. (March 2020, date last accessed).

Bouckaert RR, Frank E. Evaluating the replicability of significance tests for comparing learning algorithms.. In: *Advances in knowledge discovery and data mining. PAKDD 2004. Lecture Notes in Computer Science, Vol. 3056*, Berlin, Heidelberg: Springer, 2004.

Campana SE. Chemistry and composition of fish otoliths: pathways, mechanisms and applications. *Mar Ecol Prog Ser* 1999;**188**:263–97.

Campana SE. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. *J Fish Biol* 2001;**59**:197–242. https://doi.org/10.1111/j.1095-8649.2001.tb00127.x.

Carbonara P, Follesa MC. Handbook on fish age determination: a Mediterranean experience. In: *General Fisheries Commission for the Mediterranean. Studies and Reviews, Vol. 98*, Rome: FAO, 2019, 1–179.

Chollet F *et al.* Keras. *Github repository.* 2015. https://github.com/fchollet/keras(November 2020, date last accessed).

Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998;**10**:1895–923.

Dutta A, Zisserman A. The VIA annotation software for images, audio and video. In: *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*.New York, USA: ACM, 2019. https://doi.org/10.1145/3343031.3350535.

Fablet R, Cao F. Automatic morphological detection of otolith nucleus. *Pattern Recognit Lett* 2006;**27**:658–66.

Fablet R, Le Josse N. Automated fish age estimation from otolith images using statistical learning. *Fish Res* 2005;**72**:279–90. https://doi.org/10.1016/j.fishres.2004.10.008.

Fisher M, Hunter E. Digital imaging techniques in otolith data capture, analysis and interpretation. *Mar Ecol Progress Series* 2018;**598**:213–31. https://doi.org/10.3354/meps12531.

Formella A, Vázquez JM, Carrión P *et al.* Age reading of cod otoliths based on image morphing, filtering and fourier analysis. In: *Proceedings of the 7th IASTED International Conference on Visualization, Imaging, and Image Processing*. Anaheim: ACTA Press, 2007.

Harbitz A. A generic ad-hoc algorithm for automatic nucleus detection from the otolith contour. In: *4th International Otolith Symposium*. Monterey, USA, 2009.

He K, Gkioxari G, Dollár P *et al.* Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988, 2017.

Henderson T. correctR: corrected test statistics for comparing machine learning models on correlated samples. R package version 0.1.3. https://CRAN.R-project.org/package=correctR(last accessed May 2023). 2023.

ICES. *Report of the Workshop on Age Reading of North Sea Cod (WKARNSC)*, 5-7 August 2008, Hirsthals, Denmark. ICES CM 2008/ACOM:39. 71pp. 2008. https://doi.org/10.17895/ices.pub.19280396.

ICES. Report of the spring 2019 western Baltic cod (*Gadus morhua*) age reading exchange—SD 22. 2020, https://smartdots.ices.dk/sampleImages/2019/201/report_smartdot_event201.pdf (October 2023, date last accessed).

Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: F Pereira, CJ Burges, L Bottou, KQ Weinberger (eds.), *Advances in Neural Information Processing Systems, Vol. 25*, Curran Associates, Inc, 2012.

Krumme U, Stötera S, McQueen K *et al.* Age validation of age 0-3 cod *Gadus morhua* in the western Baltic Sea through mark-recapture and tetracycline marking of otoliths. *Mar Ecol Prog Ser* 2020;**645**:141–58.

Lin TY, Maire M, Belongie S *et al.* Microsoft coco: common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference*, September 6-12, 2014, Proceedings, Part V 13(pp.740–55). Zurich: Springer International Publishing, 2014.

Martinsen I, Harbitz A, Bianchi FM. Age prediction by deep learning applied to Greenland halibut (*Reinhardtius hippoglossoides*) otolith images. *PLoS One* 2022;**17**:e0277244. https://doi.org/10.1371/journal.pone.0277244.

Matterport. https://github.com/matterport/Mask RCNN. 2017. (November 2020, date last accessed).

Moen E, Handegard NO, Allken V *et al.* Automatic interpretation of otoliths using deep learning. *PLoS One* 2018;**13**:e0204713. https://doi.org/10.1371/journal.pone.0204713.

Moen E, Vabø R, Smoliński S *et al.* Age interpretation of cod otoliths using deep learning. *Ecol Inform* 2023;**78**:102325. https://doi.org/10.1016/j.ecoinf.2023.102325.

Nadeau C, Bengio Y. Inference for the generalization error. *Mach Learn* 2003;**52**:239–81. https://doi.org/10.1023/A:1024068626366.

Ordoñez A, Eikvil L, Salberg AB *et al.* Explaining decisions of deep neural networks used for fish age prediction. *PLoS One* 2020;**15**:e.0235013. https://doi.org/10.1371/journal.pone.0235013.

Ordoñez A, Eikvil L, Salberg AB *et al.* Automatic fish age determination across different otolith image labs using domain adaptation. *Fishes* 2022;**7**:71. https://doi.org/10.3390/fishes7020071.

O'Shea K, Nash R. *An Introduction to Convolutional Neural Networks*. CoRR, abs/1511.08458. 2015.

Panfili J, de Pontual H, Troadec H *et al. Manual of Fish Sclerochronology*. Ifremer-IRD coedition. https://archimer.ifremer.fr/doc/00017/12801/9742.pdf (October 2023, date last accessed). 2002.

Politikos DV, Petasis G, Chatzispyrou A *et al.* Automating fish age estimation combining otolith images and deep learning: the role of multitask learning. *Fish Res* 2021;**242**:106033. https://doi.org/10.1016/j.fishres.2021.106033.

R Core Team. *R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org (May 2023, date last accessed). 2020.

Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI). Lecture Notes in Computer Science, Vol. 9351*, pp. 234–41. Cham: Springer, 2015.

Sigurðardóttir AR, Sverrisson Þ, Jónsdóttir A *et al.* Otolith age determination with a simple computer vision based few-shot learning method. *Ecol Inform* 2023;**76**:102046. https://doi.org/10.1016/j.ecoinf.2023.102046.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations*. Computational and Biological Learning Society. pp. 1–14, 2015.

Szegedy C, Vanhoucke V, Ioffe S *et al. Rethinking the Inception Architecture for Computer Vision*. CoRR, abs/1512.00567. 2015.

Troadec H. Frequency demodulation on otolith numerical images for the automation of fish age estimation. *Aquat Living Resour* 1991;**4**:207–19.

Van der Walt S, Schönberger JL, Nunez-Iglesias J *et al.* scikit-image: image processing in Python. *PeerJ* 2014;**2**:e453. https://doi.org/10.7717/peerj.453.

Van Rossum G, Drake FL. *Python 3 Reference Manual*, Scotts Valley, CA: CreateSpace, 2009.

Williams AJ, Davies CR, Mapstone BD. Variations in the periodicity and timing of increment formation in red throat emperor (*Lethrinus miniatus*) otoliths. *Mar Freshw Res* 2005;**56**:529–38.

Zimmermann RS, Siems JN. Faster training of mask r-cnn by focusing on instance boundaries. *Comput Vis Image Underst*, 2019;**188**:102795.

Handling Editor: Francis Juanes