*ICES Journal of Marine Science*, 2023, **0**, 1–11
DOI: 10.1093/icesjms/fsad170
**Original Article**



# Annotating otoliths with a deep generative model

**Troels Arnfred Bojesen[1], Côme Denechaud [2], and Ketil Malde [1,2,*]**

[1]Department of Informatics, University of Bergen, Bergen 5007, Norway
[2]Institute of Marine Research, Bergen 5005, Norway

[*] *Corresponding author: tel: +47 98691834;* e-mail: ketil.malde@imr.no.

Otoliths are a central information source for fish ecology and stock management, conveying important data about age and other life history for individual fish. Traditionally, interpretation of otoliths has required skilled expert readers, but recently *deep learning* classification and regression models have been trained to extract fish age from images of otoliths from a variety of species. Despite high accuracy in many cases, the adoption of such models in fisheries management has been slow. One reason may be that the underlying mechanisms the model uses to derive its results from the data are opaque, and this lack of legibility makes it challenging to build sufficient trust in the results. Here, we implement a deep learning model that instead of age predicts the location of annotation marks for each of the annuli. This allows an expert to evaluate the model's performance in detail. The quality of the annotations was judged by a panel of four expert otolith readers in a double-blinded randomized survey. Using a scale from 1 to 5, the generated marks received an average quality score of 4.22, whereas expert annotations received an average score of 4.33. By counting the marks to determine fish age, we obtained an agreement between expert and model annotations of 64% on our test set, which running the model stochastically increased to 69%. Stochastic sampling yields further benefits, including an explicit measure of the model's uncertainty, the *post hoc* likelihood of the different age classes for each otolith, and a set of alternative annotation sequences that highlight the structure of the annuli.

**Keywords:** deep learning, explainable AI, fish age estimation.

## Introduction

The study and assessment of fish stock dynamics rely on knowledge of population age structure to track year classes through time and assess population growth and mortality, in particular for harvested populations where the effects of selective fishing must be monitored (Hidalgo *et al.*, 2011; Brunel and Piet, 2013). Because calcified structures such as otoliths and scales record physiological and environmental changes through variations in deposited material, they form temporally resolved growth zones at different time scales (Wright *et al.*, 2002). For a vast number of species, those zones can then be interpreted as individual estimates of fish age, resulting in millions of otoliths collected worldwide every year, primarily for ageing purposes (Campana, 2001; Morales-Nin and Geffen, 2015).

Despite the straightforward relationship between otolith growth rings and fish age, providing reliable estimates is time and resource intensive. Fish otoliths differ in shape and their growth rings in seasonality between fish families and species, or even at the inter-population level (Campana, 2005; Stransky *et al.*, 2008; Cadrin *et al.*, 2014). As a result, age reader training can last several years before the age estimates are reliable (Carbonara and Follesa, 2019), and uncertainties in both inter-reader agreement and "true" age accuracy will have a direct influence on the quality of the assessment.

### Deep learning for otolith classification and explainable AI

In the last two decades, new technologies and methods have therefore been explored in order to automatize, scale, and standardize otolith age reading and improve the reliability of stock assessments, among which machine learning (ML) frameworks have shown promising results. In particular, deep learning models have been shown to be effective at extracting age from Greenland halibut (Moen *et al.*, 2018), Northeast Arctic cod (Moen *et al.*, in review), red mullet (Politikos *et al.*, 2021), and also from salmon scales (Vabø *et al.*, 2021). However, these systems operate as *black box* models, and while they provide a quantitative answer, how they arrive at this answer is opaque. This lack of legibility may make users less inclined to trust the models. There exist methods that use the internal state of the model to identify parts of the input that contribute more to the result (Lipton, 2017), and such methods have been applied to otolith analysis (Ordoñez *et al.*, 2022). However, these approaches have been shown to often be unstable (Ghorbani *et al.*, 2019), the methods remain difficult to understand, and their output is not necessarily easy to interpret. It is therefore questionable whether this information makes users more willing to trust the model in practice.

### A generative model for annotations

Instead of attempting to unveil the workings of a black box classifier through *post hoc* analysis, our goal here was to generate output of a type that can be verified directly by a human expert. For otoliths, this means generating a set of annotations that mark the locations of individual annuli (see Figure 1). A user can inspect the annotations and verify that they match with the visible structures in the otolith, and the age can trivially (and automatically) be obtained by counting them. As there is considerable freedom in the placement of annotations along each annulus, standard methods for landmark or object detection cannot be used directly. Instead, we designed a process that mimics the annotation process used by expert annotators. The model first estimates the likelihood for each

**Received:** 6 July 2023; **Revised:** 17 August 2023; **Accepted:** 20 September 2023

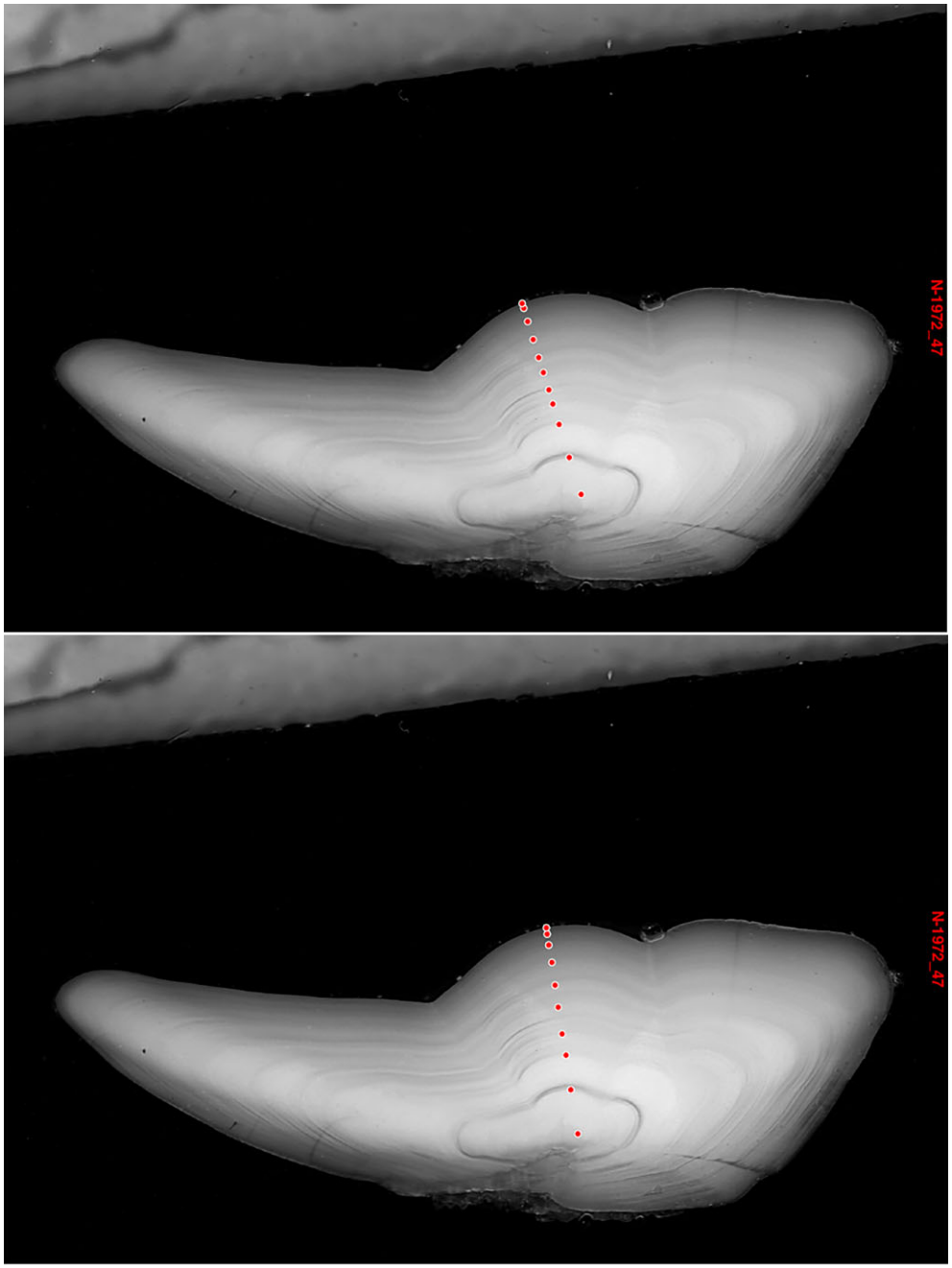


**Figure 1.** An otolith from our test set with computer generated annotations (top) and expert-generated annotations (bottom). Although the generated annotations are sometimes placed along a slightly different axis to the manual annotations, they still follow the correct otolith structure and result in the same predicted age.

pixel to be the location of the first annotation mark (which indicates the centre of the otolith). It then iteratively estimates a similar spatial probability distribution or "probability map" for the next mark conditioned on previous marks, and selects its location. In each step, it simultaneously estimates the probability for ending the sequence and terminating the iteration.

In contrast to a traditional "black box" regressor or classifier, our system produces an explanation of its estimate in terms of visual and easily interpretable information. In addition to providing improved legibility, the location of the annotation marks can be used to derive growth rates and other life history information. Furthermore, our system can estimate uncertainty as an explicit probability for each possible age of an otolith, and produce a map of the salient features detected, further boosting user confidence in the model.

## Methods

### Otolith image data

This study utilized a large dataset of adult Atlantic cod otoliths that had been sectioned, imaged, and annotated to get age and yearly growth estimates (the detailed methodology can be found in Denechaud *et al.*, 2020). A total of 4095 mature cod otoliths (ages 7+) were processed and annotated using an ObjectJ plugin (Denechaud *et al.*, 2018) that

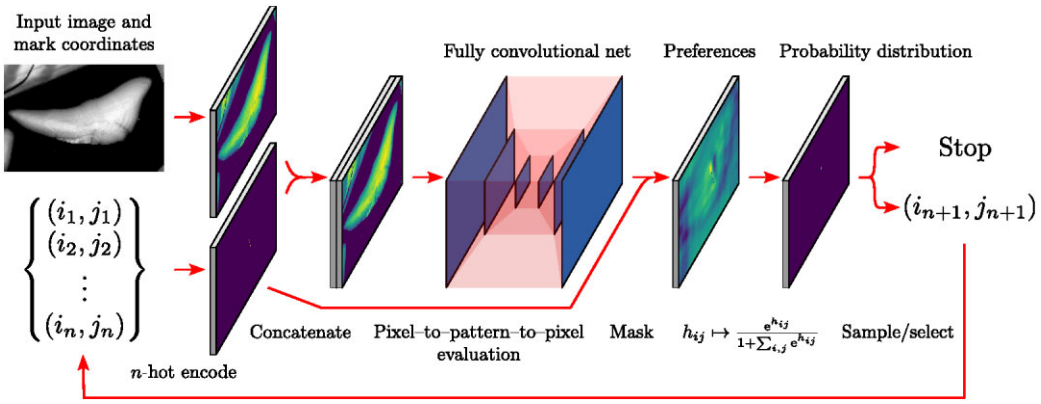


**Figure 2.** Schematic illustration of the generative model. See supplementary materials for a detailed description of the fully convolutional "pixel-to-pattern-to-pixel" network. Note that U-Net, as commonly used for semantic segmentation, outputs a set of maps for each output class, and applies a softmax function so that the probabilities for all classes sum to one for each pixel. In contrast, our model outputs a single map of probabilities, and applies the softmax over all pixels plus the stop symbol. By constraining the possible mark placements to the pixel positions of the $W \times H$ otolith image, we have thus mapped the annotation problem to that of drawing from a $(WH + 1)$-dimensional categorical distribution conditioned on the input data.

leaves nondestructive annotations stored in dedicated files. The annotations are only linked to each otolith image and as such are entirely non destructive, and are here imported using only pixel coordinates. Details of the plugin's functionalities can be found within the dedicated tutorial and the Supplementary Figure S1.

Both the unmodified grayscale images and the associated annotations were imported and coupled by individual ID, then processed for efficient treatment in the neural network model. All images had been consistently annotated on the distal axis (shortest growth axis from the core), which ensured the model would only attempt annotations in this general direction rather than haphazardly across the otolith surface.

First, the otolith images were cropped and downscaled to a resolution of 1200 by 800 pixels to reduce memory footprint and improve training times. The corresponding annotation coordinates were transformed similarly and rounded to the nearest integer values. In four cases where the rings at the edge were narrow and the annotations were close to each other, this rounding caused the transformed annotations to have identical coordinates. As our model presumes that all marks should be spatially distinguishable from each other, we discarded these data points.

## Conditional probability neural network model

A schematic illustration of the model for generating sequences of annotation marks is shown in Figure 2. The model generates marks, specified by their pixel coordinates *(i, j)*, iteratively by mapping an image and any previously generated marks into a probability map, from which it can be determined whether and where a new mark should be placed. In other words, this is an autoregressive generative model.

The model performs the following steps:

1. First, the model maps the previously generated marks into an $n$-hot encoded "image," with pixels set to be one where there is a mark and zero everywhere else. The $n$-hot encoding renders the coordinate information in the same form as the otolith image, making it possible to pass a concatenation of the two arrays into a fully convolutional "pixel-to-pattern-to-pixel" network.
2. The concatenated arrays are processed using an architecture similar to U-Net (Ronneberger *et al.*, 2015), and are described in more detail in supplementary materials. The network outputs a single channel map for the placement of the next mark, and this map is then masked by the $n$-hot encoding of the previous marks to produce the "preference image," $h$. The masking consists of subtracting a large number for each pixel where there is a previously placed mark, effectively making the probability of subsequently selecting these pixels zero.
3. Finally, we convert the preferences into probabilities using a softmax, i.e. we assign a probability weight of 1 to the selection of the "end sequence" action and a probability weight of $exp(h_{i,j})$ to the selection of pixel *(i, j)* for the placement of the next mark, and normalize.

## Training procedure

The annotated image data were split into training, validation, and test sets consisting of 3591, 200, and 300 annotated images, respectively.

To train the model, we randomly select the $n \in \{0, 1, \ldots l\}$ first marks of a human-annotated otolith with $l$ marks. As described in the section above, the model is given the otolith image and the mask representing these first $n$ marks, and outputs a probability map for the placement of the $n + 1$'th mark. Using this map, we minimize the negative log-likelihood of placing the $n + 1$'th mark in the annotated location, or for ending the sequence if $n = l$.

We trained the model on a single NVIDIA A100 GPU with a batch size of 32. We used the Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of $10^{-3}$ for 100 epochs, followed by 20 epochs of fine-tuning with a learning rate of $10^{-4}$. The decay rates were kept at the default 0.9 and 0.999 for the first and second order momentum estimates, respectively. In this context, we denote an epoch to be the processing of each training set otolith image once. Note, however, that for each image, a random length input mark sequence is drawn and a random flip augmentation about the short axis of the image is used. The former means that each image with human annotations constitutes $\sim 10$ individual training pairs (an input paired with its expected output), while the latter ef-

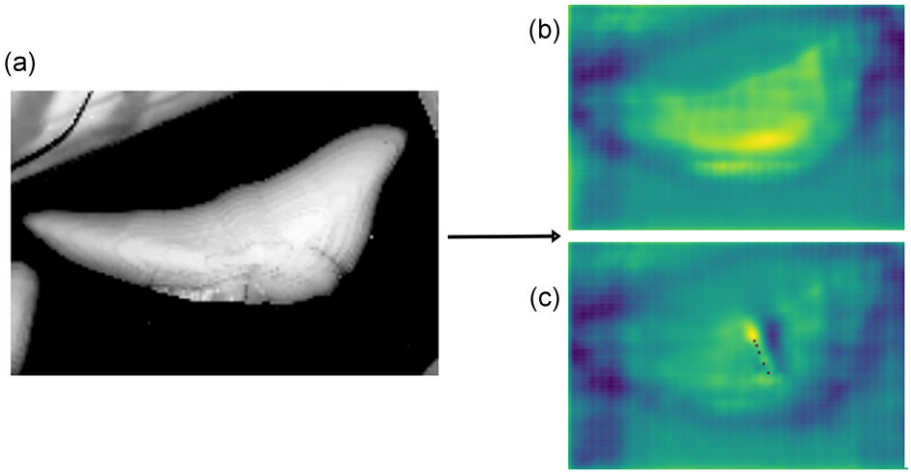


**Figure 3.** Example of annotation probability maps for an otolith image (a) with either zero (b) or five (c) previously determined annotation marks. For clarity, the probabilities are shown log-transformed, with warmer colours indicating higher probabilities. Note that for (c), we can see the preceding five marks visible as black dots, as the masking process sets their probability to zero. As (b) is determining the placement of the first mark, no such masking is done. For illustrative purposes, the resolution has been reduced to 150 × 100 pixels.

fectively doubles the available training set. Overall, this means that each training pair was encountered only ∼ 6 times during training.

### Generating annotations

When the model has been trained, we can use it to generate annotations for hitherto unseen and unannotated otolith images. We either do this stochastically or greedily. In the stochastic approach, we generate a sequence of annotation marks by sampling from the conditional probability map the model outputs. In practice, we do this by using the Gumbel-max trick (Jang *et al.*, 2017), which allows us to sample directly using the preferences, rather than having to map to the probability distribution first. In the greedy approach, we deterministically generate a "most likely" mark sequence as follows: For each iteration, we first determine whether the probability of ending the sequence (according to the model) is larger or smaller than 0.5. If it is larger, we end the sequence. If it is smaller, we place a mark at the pixel coordinate with the largest preference, $i_{n+1}\, j_{n+1} = argmax_{i,j}$. Intuitively, we do it in this way because spatial placement and termination selection probabilities have to be treated differently when we are selecting the most likely next action. The mathematical rationale is explained in the supplementary materials.

Because the annotation sequence is characterized by a starting point (placed at the centre of the otolith core) and an ending point (placed at the edge of the otolith), estimates of fish age can be calculated by subtracting 2 from the sequence length of a given sample. Figure 3 shows an example probability map for an otolith with zero and five previously determined annotation marks. We see that the spatial uncertainty for where to place the next mark is larger in the case of placing the first as compared to the sixth mark.

### Evaluation by expert readers

In order to evaluate the quality and credibility of generated annotations, four expert readers were asked to score a randomized sample of manually annotated images and their automatically generated counterpart in a double-blind experiment. A total of 120 annotated images (60 otoliths) were randomly selected from the test set and partitioned into 4 sets of 30 images that were manually adjusted to ensure each sample contained exactly 15 images annotated by an expert and 15 images generated by the network. Each reader was given a sample and asked to grade each annotation on a scale from 1 to 5, with 5 being the best (see Table 1). No reader was offered both computer and expert annotated images from the same otolith, and readers were asked to provide relevant comments related to the scoring decision (such as clear rings left unannotated, false zones being mistakenly annotated, etc).

**Table 1.** Grading scale for evaluating annotation quality.

| | |
|---|---|
| 5 | Close to perfect annotation |
| 4 | Minor inaccuracies of little consequence for age estimate |
| 3 | Some errors not expected from an expert annotator |
| 2 | Low quality annotation with major mistakes |
| 1 | Unusable annotation, not useful at all |

## Results

### Accuracy of predicted ages

Since the age classes were quite imbalanced, we investigated accuracy for the most abundant age classes (8, 9, and 10 years, corresponding to 10, 11, and 12 marks) individually (Table 2). Since the true age is not known, accuracy here refers to the proportion of predicted ages that match the manually annotated age, where the human estimates are treated as ground truth. For the two most abundant classes (8 and 9 years), we see that the accuracy is consistent at close to 0.7, with the stochastic method consistently better than the greedy method. For most of the metrics, we see a noticeable improvement when applying the stochastic method. The main exceptions are mean absolute error (MAE) and root mean squared error (RMSE) for 9 year old where the larger number of zero-length sequences leads to inflated error values (see the zero column of Figure 4 and rightmost bars in Figure 5).

### Monte Carlo sampling the stochastic method

We ran the model multiple times, sampling each annotation mark according to its conditional probability, and accumulat-

**Table 2.** Accuracy, MAE, and RMSE for the most abundant age classes in the training set.

| Age class | N (test set) | Greedy accuracy | Greedy MAE | Greedy RMSE | Stochastic accuracy | Stochastic MAE | Stochastic RMSE |
|---|---|---|---|---|---|---|---|
| 8 | 220 | 0.68 | 0.4 | 1.0 | 0.72 | 0.4 | 1.1 |
| 9 | 43 | 0.67 | 0.3 | 0.6 | 0.70 | 0.5 | 1.8 |
| 10 | 30 | 0.37 | 0.9 | 1.2 | 0.53 | 0.6 | 0.9 |

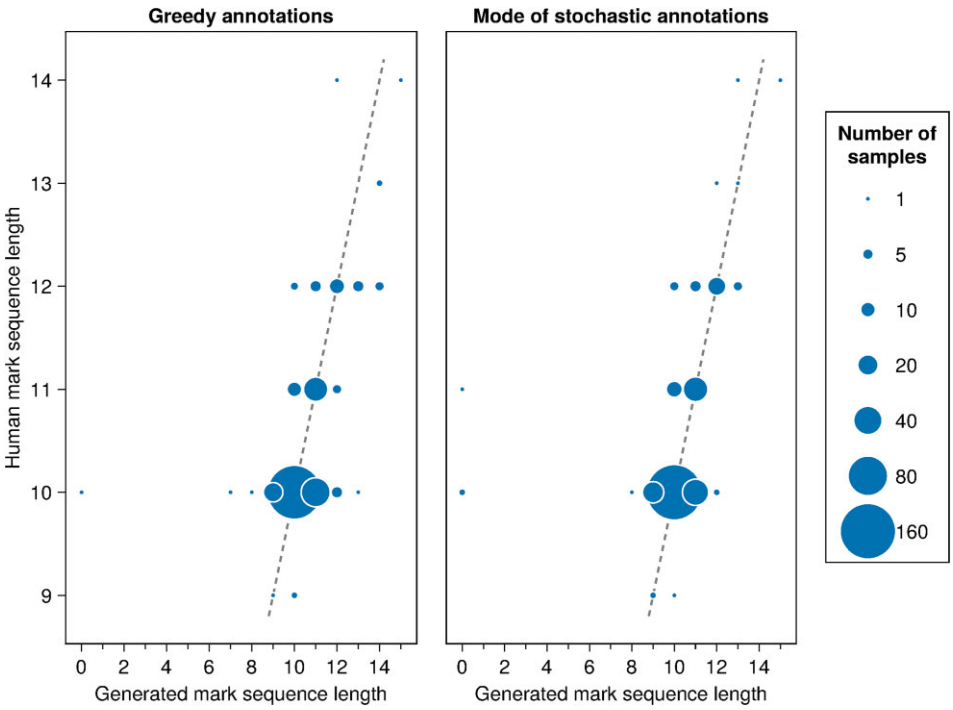


**Figure 4.** Test set distribution of generated sequence lengths organized according to their human expert mark sequence length. The area of a circle is proportional to the number of samples in the given age category. The dashed lines indicate a perfect correspondence between computer generated and human expert sequence lengths.

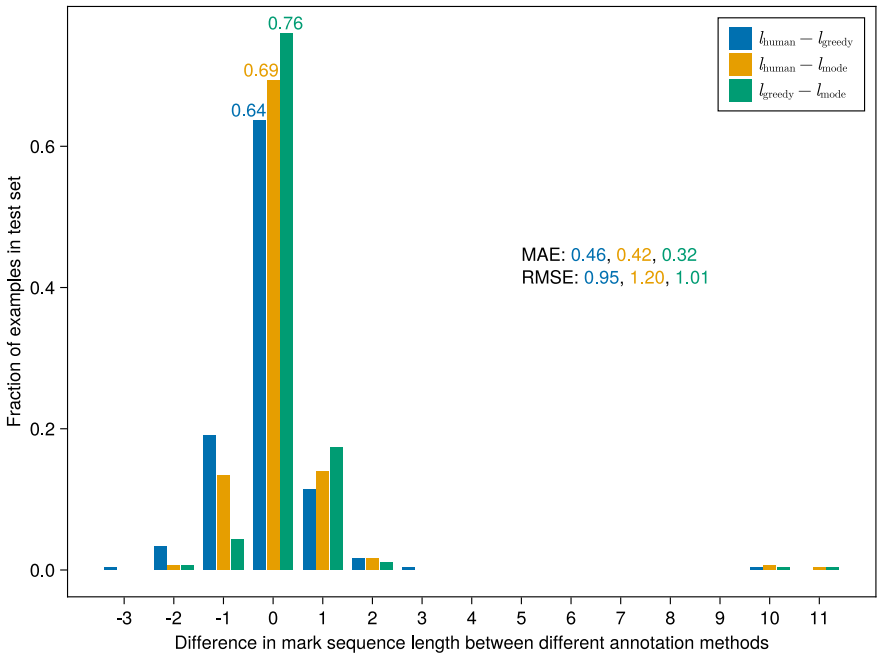


**Figure 5.** Test set distribution of the differences in annotation sequence lengths (corresponding to predicted age plus two) between the human expert, $l_{human}$, the greedy algorithm, $l_{greedy}$, and the mode of the stochastic algorithm, $l_{mode}$. The accuracy of the greedy and mode approaches is 0.64 and 0.69, respectively. MAE and RMSE refer to the mean absolute error and the root mean squared error, where "error" in this context is the difference in number of annotation marks.

ing the probability distributions for each mark (see supplementary materials for details). This achieved two objectives, (i) it allowed us to visualize the salient features of the otolith (*viz.*, the putative annuli as seen by the model), and (ii) it let us derive a probability distribution for the age estimates.

Two examples based on Monte Carlo simulations of 1024 independent sequences are shown in Figure 6. We notice that in both cases, the starting point differs from the labelled start point, and in the right image, this leads to the model following a markedly different axis. This is not incorrect *per se*, but may indicate more ambiguity in the image. In general, the model consistently identifies and annotates ring-like structures in line with those from human experts, though the Monte Carlo resampling shows some level of lateral spread along the rings based on the origin and direction of the iterative annotation process.

### Qualitative evaluation experiment

Human and computer generated annotations performed similarly well when assessed by expert age readers, although human annotations had on average slightly higher scores (Figure 7). Reader B had the largest difference in scoring (4.27 versus 3.73) while reader C was the only one with a higher mean score attributed to computer generated annotations (4.53 versus 4.80). Across all expert readers, human annotations had a mean score of 4.33 and computer generated annotations had a mean score of 4.22 (Figure 7).

Poor quality annotations were identified in both the human and computer generated sets, although more were attributed a score of 1 or 2 in the computer generated set (indicating unusable annotations with major mistakes, cf. Table 1). An identical number of 39 images had the highest score of 5 in both sets, and few images with a high score (4 or 5) in one annotation category had a low score in the other. Those cases corresponded mostly to otoliths where too many or too few zones were annotated in one of the sets, such as the computer annotating extra rings from visual artefacts found outside the otolith, or the original human reader hopping over a ring considered to be part of a split zone (Figure 8).

Out of the 60 otoliths used for the inter-reader test, 36 had the same predicted age in both the human and computer generated annotations and the rest differed by one year (21 otoliths), two years (2 otoliths) and three years (1 otolith). Otoliths with identical ages generally corresponded to clear images with high scores (4 or 5) in both sets, with the exception of 6 otoliths where ages differed by one year yet the test readers evaluated their respective image as correct (Table 3).

## Discussion

This study presented a first exploration of generated otolith annotations as an alternative to traditional ML regressions for age reading. By conditioning the algorithm to identify and annotate ring structures as read by human experts, we show the potential for more transparent and applicable methods that move away from "black box" regressions.

### Accuracy of age estimates from generated annotations

This study found a 64% overall agreement between expert age estimates and generated annotations for the greedy approach and 69% for the stochastic approach, which compared favourably to other methods recently published. Sigurðardóttir *et al.* (2023) summarized results from several recent applications of deep neural networks to otolith age estimations, where reported accuracies varied from 0.2 to 0.69. It should be clear that accuracies across data sets are not comparable, and depend to a large degree on the range of ages, the quality of images and labelling, and the quantity of training data. Using the same approach on several data sets produced accuracies that varied from 0.30 on Greenland halibut to 0.61 on Haddock. Our results must therefore be seen in the context of the limited age range in our data and very-high quality images and annotations. Additionally, while 64% overall agreement is still lower than the desired output from inter-reader benchmarks (which usually aim for at least 80% agreement), more than 90% of the dataset fell within either perfect agreement or generated annotations that differed from human experts by only 1 increment (either in deficit or surplus). This indicates that the algorithm was successful in replicating human annotation processes with remarkable precision, and that further training and fine-tuning with a larger sample size and class balance may well increase agreement close to desired levels.

One weakness with traditional classifiers is the detrimental effect of imbalanced class abundances (Johnson and Khoshgoftaar, 2019). Otoliths, like many other data sets, often have uneven age distributions, and classifiers will typically have lower performance for scarce classes. This has also been observed for regressors (Moen *et al.*, 2018) and for classifiers using new self-supervised methods (Sigurðardóttir *et al.*, 2023). Here, we train on each annotation mark individually, and thus the model has no explicit concept of age classes, or even age. Although there are more 10-mark sequences than 9-mark sequences in our data (due to the abundance of 8 years old in the dataset), counting is done separately from the model, eliminating much of the bias.

In a few cases, the model outputs a sequence of length zero, in other words, it terminates the sequence before setting any mark. This is detrimental to the statistics, in particular measurements of the magnitude of errors, like RMSE. Although this is clearly an incorrect result, this information can be used to identify troublesome otoliths for evaluation by human experts. Granted some stronger assumptions based on knowledge of the age distribution in the training data, this issue could ultimately be trivially mitigated by implementing a minimum number of steps before the sequence is allowed to terminate. Similarly, the distribution of age predictions resulting from sampling the stochastic method can be used to estimate probabilities for the various age classes, and to generate different proposals for annotation mark placements for an expert user to evaluate. Such corrections and their effect on the prediction accuracy could be explored in future work granted a balanced training set including all ages from the youngest to the oldest.

### Advantages of an iterative rather than a one-shot approach

In an earlier attempt at generating annotations for otoliths, we employed a GAN (Goodfellow *et al.*, 2020) to generate all annotations in a single pass. Our current, autoregressive approach comes with several advantages over such a one-shot approach. First, dividing the problem into a sequence of similar sub-problems greatly simplifies the output complexity the

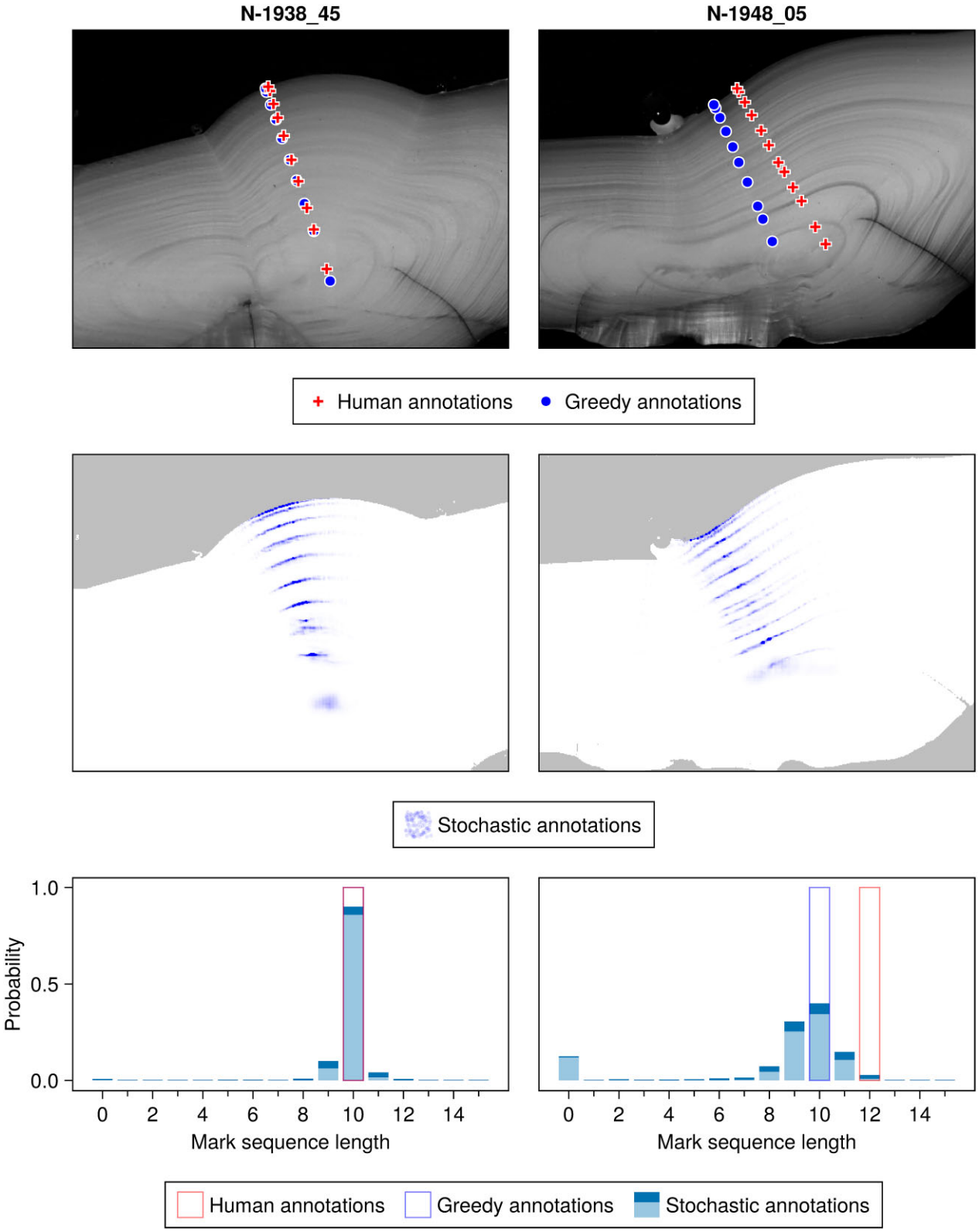


**Figure 6.** Examples of stochastic generation of annotations for a clear (left column) and a more uncertain (right column) case, compared with the respective corresponding human and greedy annotations (top row). The middle row shows the spatial probability distribution for generating a mark (stronger blue means higher probability). The target otolith image and its silhouette are visible in the respective backgrounds. The bottom row illustrates the distribution of mark counts. The darker bands at the top of the bars illustrate 95% confidence intervals, as found by bootstrapping.

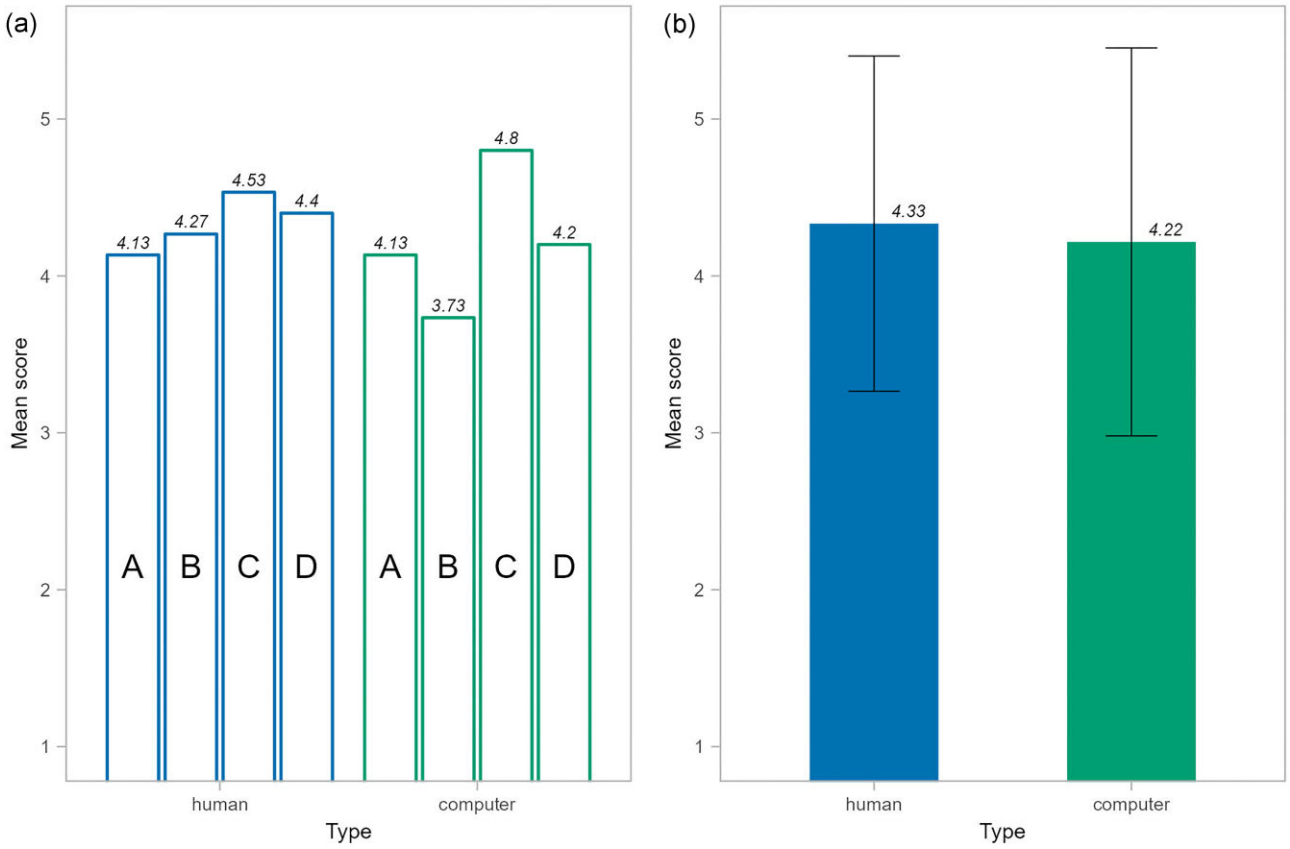


**Figure 7.** Inter-reader (a) and general (b) comparison of mean quality scores given to human and computer generated annotations. Letters identify individual readers.

neural network has to learn. In our case, by making the target a single point in pixel-space, the task is reduced to that of a standard (albeit high-dimensional) classification problem. Second, by taking the annotation order implicitly present in the training data into account, the generative approach more closely follows the human way of annotating. This approach also effectively enlarges the training data set by a factor proportional to the average number of annotation marks for an image. In other words, each annotated image now constitutes several training data points. It is also noteworthy that, in aggregate, these qualities make our autoregressive approach fairly lightweight. The model only took a few hours to train and is sufficiently lightweight to be trainable on a laptop GPU, although limited GPU memory may restrict the batch size.

## Value of explicit annotations

When evaluated by expert cod readers in a blind experiment, our generated annotations were well-received and generally unidentifiable from human annotations. The scoring approach highlighted the fact that annotation quality and usability were directly dependent on the otolith image quality rather than the source of the annotations. When images were of lower quality (typically, too dark and not enough contrast between zones), they were consistently scored worse no matter who annotated them. As pointed out in previous works on the use of deep-learning for ageing otoliths, this confirms the importance of standardized image taking procedures to ensure the highest quality training data with a minimal amount of "extrinsic noise" (Fisher and Hunter, 2018; Martinsen *et al.*, 2022).

When the pattern of zones on the otolith itself was unclear, both human and generated annotations tended to be attributed lower scores, or alternatively a strong mismatch between the original annotations and their generated counterpart was visible. These unclear or difficult otolith patterns may emerge in periods of disturbed growth rates caused by detrimental conditions such as suboptimal temperatures or food availability and decreased seasonality (Fowler, 1995; Høie *et al.*, 2009; Albuquerque *et al.*, 2019). While age reading of cod and most heavily fished species is well understood and grounded in age validation experiments (Campana, 2005), there is an unavoidable level of subjectivity in human estimates that comes with experience. These small differences will manifest in particular when an otolith is unclear and has room for personal interpretation, for example when the innermost ring is hard to separate from the settlement ring, or when a given individual has split, unclear zones that a human reader will selectively annotate or ignore based on fish length, catch date, and other variables. The present algorithm, on the other hand, will be operating in a more standardized manner and may be more likely to annotate all detectable ring structures that match the growth pattern most commonly seen in the training data, instead of selectively hopping over zones that seem split or unclear.

Altogether, this indicates that the algorithm was successfully trained to identify rings and structures the way human

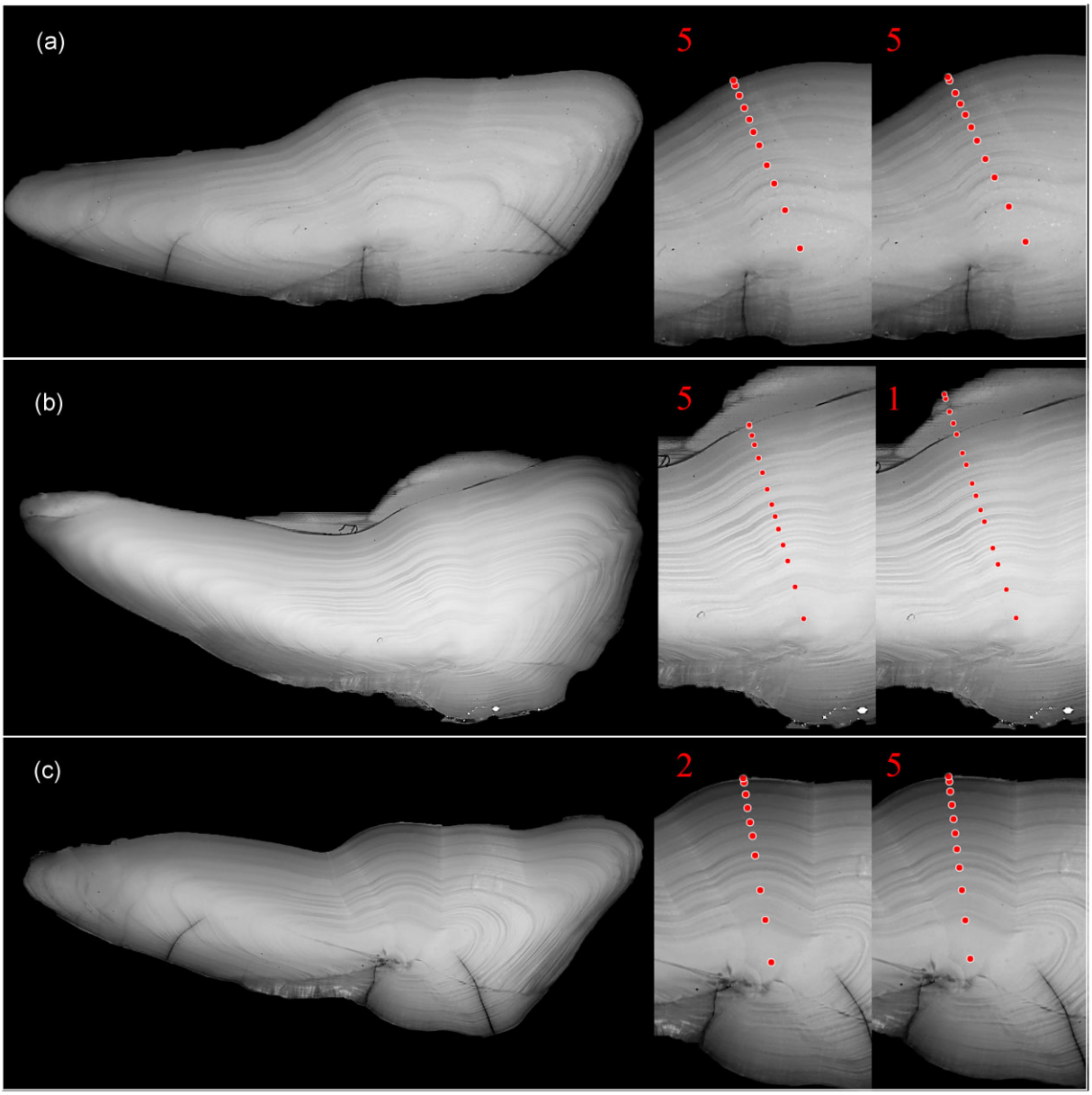


**Figure 8.** Otoliths examples and their respective human (left) and computer generated (right) annotations in three scenarios: perfect score and agreement in both (a); high human score but poor computer score (b); and poor human score but high computer score (c). Scores are given in red for each annotation set.

**Table 3.** Number of human and computer generated annotations within each scoring category for the inter-reader test.

| Score | Human | Computer |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 5 | 8 |
| 3 | 6 | 4 |
| 4 | 9 | 7 |
| 5 | 39 | 39 |
| Total | 60 | 60 |

readers do. This opens up new windows of application for ML assisted frameworks. Because the machine is solely trained to identify and annotate rings, the actual age estimation is left to readers to assess. This addresses some of the common issues with the applicability of regression methods, namely that the provided estimate often comes from a "black box" with little to no indication of what kind of information was used to derive this result. Such generated annotations could have a wide range of applications, from introducing external estimates in reading workshops to investigate reader-specific trends; using generated annotations as training material; or multiplying the training data available for more complex regression models by consistently generating annotations and validating them with human readers. Because the majority of the time and resources needed to age read fish come from collecting, extracting, preparing, and imaging the otoliths prior to reading (Fisher and Hunter, 2018), ML approaches are usually limited in scope by how efficient they actually are: generating annotations from already available material could potentially be used to multiply the amount of training data available for more other models without having to use additional laboratory and reader time.

### Trust, interpretability, and explainable AI

Despite remarkable progress in AI and deep learning technology over the last decade, adoption in the marine sciences and elsewhere is slow (Thessen, 2016; Lopez-Marcano *et al.*, 2021). One possible reason for this may be that users do not trust complex models whose operation is difficult to decipher. This opacity is widely acknowledged as a challenge for deep neural networks, and may lead analysts to prefer models with simpler computational structures (e.g. rule based models, decision trees, or linear regression) instead. However, we contend that this perceived simplicity is a consequence of few parameters, rather than inherent in the algorithms. To match the decision power of a deep neural network, millions of rules, tree nodes, or dimensions are needed, at which point the method is no longer explainable or interpretable.

An alternative is to apply post hoc analysis to produce "explanations" (Doran *et al.*, 2017), usually in the form of highlighting the parts of the input that contribute most to the result. The post hoc analysis is often itself opaque, and thus suffers from the same problem as it is attempting to address. Even in the case where the analysis identifies features the user considers salient, one might reasonably ask whether this proves anything about the decision process.

What constitutes an "explanation" of a result or "interpretability" of a model is not clear (Lipton, 2017). We consider an explanation to be auxiliary information that the user can use to convince himself of the correctness of a result. With Monte Carlo sampling, our method produces salience maps similar to *post hoc* analyses, but our "explanations" are generated as part of the algorithms and used explicitly by it to determine age. How the probability maps for each annotation are produced remains opaque, but analogous to how decomposing the system in stages simplifies its learning, it likewise simplifies the process of verifying its efficacy for the user. The link from an otolith image to probable placements of an annotation is easier to verify than the link from image to age estimate. Once this is settled, the process of age determination and calculation of probability distributions is easy to understand.

## Conclusion

Otoliths are an important tool for ecology and fish biology, but the skill required for expert analysis is limiting their use. Deep learning models can scale up the analysis, and have been shown to produce statistically adequate age estimates, but the black box nature of such systems contributes to their slow adoption. Like traditional classifiers, our model produces high-quality age estimates, but also detailed intermediate information of the process, in a form familiar to experts. Thus, the expert can readily inspect and verify the model's individual annulus identifications and correct any errors. In addition, the quality and uncertainty associated with the interpretation of each otolith can be assessed, automatically or m anually.

## Acknowledgements

The authors would like to thank the age readers and cod otolith experts at the IMR for their assistance with the reading and quality scoring of otoliths: Erlend Langhelle, Malin Skage, Eirik Oddland, and Celina Bjånes. This work was in part inspired by the Master's thesis of Emir Zamwa, from the University of Bergen (Zamwa, 2023).

## Supplementary data

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

## Author contributions

KM proposed using a generative model for otolith annotations. TB designed the algorithm and model, and was responsible for its implementation. CD provided otolith data, performed the manual annotations, and organized the expert panel for evaluating the generated annotations. All authors contributed to writing the manuscript.

## Data availability

Otolith images, predictions, and source code available on request from the authors.

## Conflict of interests

The authors declare no competing interests.

## References

Albuquerque, C. Q., Lopes, L. C. S., Jaureguizar, A. J., and Condini, M. V. 2019. The visual quality of annual growth increments in fish otoliths increases with latitude. Fisheries Research, 220: 105351.

Brunel, T., and Piet, G. J. 2013. Is age structure a relevant criterion for the health of fish stocks? ICES Journal of Marine Science, 70: 270–283.

Cadrin, S. X., Karr, L. A., and Mariani, S. 2014. Chapter one—stock identification methods: an overview. In Stock Identification Methods, 2nd edn, pp.1–5. Ed. by S. X. Cadrin, L. A. Kerr, and S. Mariani Academic Press, San Diego, http://www.sciencedirect.com/science/article/pii/B9780123970039000011 (last accessed 26 May 2020).

Campana, S. E. 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. Journal of Fish Biology, 59: 197–242.

Campana, S. E. 2005. Otolith science entering the 21st century. Marine and Freshwater Research, 56: 485–495.

Carbonara, P., and Follesa, M. C. 2019. Handbook on Fish Age Determination: A Mediterranean Experience. General fisheries commission for the mediterranean. Studies and Reviews, 98: 1–179.

Denechaud, C., Smoliński, S., Geffen, A. J., Godiksen, J. A., and Campana, S. E. 2020. A century of fish growth in relation to climate change, population dynamics and exploitation. Global Change Biology, 26: 5661–5678.

Denechaud, C., Thorsen, A., and Vischer, N. 2018. ObjectJ: Measuring Growth Rings in Fish Otoliths. https://sils.fnwi.uva.nl/bcb/objectj/examples/otoliths/MD/otoliths.html (last accessed Oct 2023)

Doran, D., Schulz, S., and Besold, T. R. 2017. October 2. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. arXiv. http://arxiv.org/abs/1710.00794 (last accessed 3 July 2023).

Fisher, M., and Hunter, E. 2018. Digital imaging techniques in otolith data capture, analysis and interpretation. Marine Ecology Progress Series, 598: 213–231.

Fowler, A. J. 1995. Annulus formation in otoliths of coral reef fish-a review. In Recent Developments in Fish Otolith Research, pp. 45–63. Ed. by D. H. Secor, J. M. Dean, and S. E. Campana University of South Carolina Press, Columbia, SC.

Ghorbani, A., Abid, A., and Zou, J. 2019. Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33: 3681–3688.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A *et al.* 2020. Generative adversarial networks. Communications of the ACM, 63: 139–144.

Hidalgo, M., Rouyer, T., Molinero, J. C., Massutí, E., Moranta, J., Guijarro, B., and Stenseth, N. C. 2011. Synergistic effects of fishing-induced demographic changes and climate variation on fish population dynamics. Marine Ecology Progress Series, 426: 1–12.

Høie, H., Millner, R. S., McCully, S., Nedreaas, K. H., Pilling, G. M., and Skadal, J. 2009. Latitudinal differences in the timing of otolith growth: a comparison between the Barents Sea and southern North Sea. Fisheries Research, 96: 319–322.

Jang, E., Gu, S., and Poole, B. 2017. August 5. Categorical Reparameterization with Gumbel-Softmax. arXiv. http://arxiv.org/abs/1611.01144 (last accessed 3 July 2023).

Johnson, J. M., and Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. Journal of Big Data, 6: 27.

Kingma, D. P., and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv. http://arxiv.org/abs/1412.6980 (last accessed 3 July 2023).

Lipton, Z. C. 2017, March 6. The Mythos of Model Interpretability. arXiv. http://arxiv.org/abs/1606.03490 (last accessed 3 July 2023).

Lopez-Marcano, S., Brown, C. J., Sievers, M., and Connolly, R. M. 2021. The slow rise of technology: computer vision techniques in fish population connectivity. Aquatic Conservation: Marine and Freshwater Ecosystems, 31: 210–217.

Martinsen, I., Harbitz, A., and Bianchi, F. M. 2022. Age prediction by deep learning applied to Greenland halibut (*Reinhardtius hippoglossoides*) otolith images. PLoS ONE, 17: e0277244.

Moen, E., Handegard, N. O., Allken, V., Albert, O. T., Harbitz, A., and Malde, K. 2018. Automatic interpretation of otoliths using deep learning. PLoS ONE, 13: e0204713.

Morales-Nin, B., and Geffen, A. J. 2015. The use of calcified tissues as tools to support management: the view from the 5th International Otolith Symposium. ICES Journal of Marine Science, 72: 2073–2078.

Ordoñez, A., Eikvil, L., Salberg, A.-B., Harbitz, A., and Elvarsson, B. Þ. 2022. Automatic fish age determination across different otolith image labs using domain adaptation. Fishes, 7: 71.

Politikos, D. V., Petasis, G., Chatzispyrou, A., Mytilineou, C., and Anastasopoulou, A. 2021. Automating fish age estimation combining otolith images and deep learning: the role of multitask learning. Fisheries Research, 242: 106033.

Ronneberger, O., Fischer, P., and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. https://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a/ (last accessed 3 July 2023).

Sigurðardóttir, A. R., Sverrisson, Þ., Jónsdóttir, A., Gudjónsdóttir, M., Elvarsson, B. Þ., and Einarsson, H. 2023. Otolith age determination with a simple computer vision based few-shot learning method. Ecological Informatics, 76: 102046.

Stransky, C., Baumann, H., Fevolden, S.-E., Harbitz, A., Høie, H., Nedreaas, K. H., Salberg, A.-B *et al.* 2008. Separation of Norwegian coastal cod and Northeast Arctic cod by outer otolith shape analysis. Fisheries Research, 90: 26–35.

Thessen, A. 2016. Adoption of machine learning techniques in ecology and earth science. One Ecosystem, 1: e8621.

Vabø, R., Moen, E., Smoliński, S., Husebø, Å., Handegard, N. O., and Malde, K. 2021. Automatic interpretation of salmon scales using deep learning. Ecological Informatics, 63: 101322.

Wright, P. J., Panfili, J., Morales-Nin, B., and Geffen, A. J. 2002. Types of calcified structures: A. Otoliths. In Manual of Fish Sclerochronology, pp. 31–57. Ed. by J. Panfili, H. d. Pontual, H. Troadec, and P. J. Wright Ifremer-IRD coedition, Brest, France.

Zamwa, E. 2023. January 30. Generative Adversarial Networks for Annotating Images of Otoliths. The University of Bergen. https://bora.uib.no/bora-xmlui/handle/11250/3060232 (last accessed 3 July 2023).

*Handling editor: Allen Andrews*



**Received:** 6 July 2023; **Revised:** 17 August 2023; **Accepted:** 20 September 2023