INTERNATIONAL PACIFIC HALIBUT COMMISSION

Using artificial intelligence (AI) for supplementing Pacific halibut age determination from collected otoliths – project update

PREPARED BY: IPHC SECRETARIAT (B. HUTNICZAK; 11 SEPTEMBER 2026)

PURPOSE

This document summarizes current knowledge on the use of artificial intelligence (AI) for determining the age of fish from images of collected otoliths and provides an update on the ongoing exploratory work to develop an AI-based age determination model for Pacific halibut. The primary objective is to assess the viability of an AI-based approach as a supplement to existing Pacific halibut ageing protocols, while also identifying the remaining steps and requirements necessary for potential operational implementation.

The progress summarized in this document focuses on:

- comparison of potential use case for AI vs. traditional break and bake (BB);
- designing mix method approach;
- assessing ageing error of a 50/50 break and bake (BB) / AI mix;
- evaluating new model architectures;
- assessing performance of surface images; and
- assessing performance of using Z-stack images (limited).

UPDATES SUMMARY

Use case comparison

At the SRB028, the SRB made the following request:

NOTING that the Secretariat has developed multiple aging methods (epigenetic, AI-based, traditional break and burn otolith aging), the SRB **RECOMMENDED** that a table be developed – with a column for each method - describing:

a. the different anticipated use cases;
b. the relative costs (split into development and production) and benefits of these approaches;
c. timelines for development; and
d. decision points that will determine future allocation of effort to this method.

	Break and bake	**AI-based ageing (CNN)**
Anticipated use cases	Primary/ reference ageing method for stock assessment; Source of validated ages used to train and benchmark AI	Supplement to manual ageing; Could be used in mix method approach (see 50/50 mix method approach proposed design)
Development costs	None (established method); real cost is training a new ager - considerable, months to proficiency	Developed in-house as a side project - intermittent staff time (decreasing with AI-assisted coding), not separately budgeted; imaging camera $700; virtual

		machine cost (model testing; 2026) $2,623
Production costs	Minimum $3.05 per otolith, for ageing only; all-in cost in-house probably closer to $4.70-$6.50 (excludes training; higher band applies typically to newer staff that are not as proficient and need more time per otolith) **Break-and-bake processing** (~$0.40/ otolith, quickly learned) is required by manual ageing and AI alike, so it does not differentiate the methods.	Break-and-bake processing shared/ identical to manual, then: **Imaging** $0.37-0.49 (~46 s/otolith; ~2,700/ week; requires ~15-min training) done by technician; IPHC is also often supported in this task by volunteers (free); workflow managed via an interactive app (minimal time); **Azure compute** ~ $300 per model build (15 runs × 38 h × ~$0.53/h); **Inference** (local; ~$0)
Benefits	Established, validated, accepted for assessment; Reference against which other methods are judged; No new infrastructure	Low marginal cost Readily scalable - imaging throughput can be increased quickly, whereas manual ageing is constrained by difficult hiring and a lengthy, specialized training process for new agers. Ages can be rerun on the same images with better models over time (improving architecture and increasing training data)
Timelines	N/A (established); new-ager training to proficiency in months	Operational at the 50/50 rate on current data; approximately one additional year recommended to verify multi-year reliability (temporal generalization).
Decision points	Ongoing reader precision monitoring	Accuracy and precision vs. BB: assessed - the 50/50 BB/AI mix showed ageing error close to BB-only (Punt et al., 2008); acceptable for implementation at the 50/50 rate on current data. Temporal generalization: additional data spanning more years required to confirm that performance holds across years before full operational reliance. However, the 50/50 mix method offers built-in adaptation to change. Ongoing evaluation of older-age performance and workflow integration.

Epigenetic ageing component is covered in IPHC-2026-SRB029-05.

Designing mix method approach

The proposed operational design integrates manual and AI-based ageing into a single production workflow (Figure 1). Its objectives are to retain assessment-quality ages while reducing cost and dependence on specialist ageing capacity, and to build in the data streams needed to quantify and monitor ageing error over time. The design is also intended to scale readily as production or capacity needs change.

All otoliths in the annual production collection are processed by break-and-bake (BB), which is required for both manual reading and AI imaging (CNN performance on surface images is currently considerably worse). The processed collection is then randomly divided into two halves:

- Manual half (50%) — aged by a single expert break-and-bake read (BB1). These reads provide production ages for half of the collection and simultaneously serve as labelled data to train or fine-tune the CNN model.
- CNN half (50%) — aged by the CNN model.

Because the CNN is updated annually as additional years of labelled data accumulate, its predictions are expected to improve over time. This depends on a portion of the manual production being photographed each year to supply that labelled data. When the model is updated, ages for the CNN-aged portion are re-estimated, so the age structure is revised retrospectively across all years rather than only going forward.

The two halves are pooled to form the age structure used in the stock assessment - 50% manual and 50% CNN ages - giving complete coverage of the production collection from a randomly assigned mix of the two methods. Random assignment ensures the halves are comparable, so the pooled age structure is not subject to selection bias between methods. [1]

Ageing error is quantified from a quality-control (QC) subset: a subsample (~10%) receives an independent second break-and-bake read (BB2), with the QC sample growing over time. These paired reads support annual estimation of the ageing-error matrix following Punt et al. (2008). Because the CNN is revised over time and its ages re-estimated for all years, the ageing-error estimates are likewise updated annually and applied retrospectively across all CNN production. Otoliths selected for QC are excluded from CNN training going forward, so QC always evaluates images the model has not already seen.

Note: Total ageing effort remains ~110% - every otolith aged once plus a ~10% QC second read - the same level as current practice. Mixed method approach decreases the manual burden from ~110% to ~60%.

[1] In practice, this would be random selection by predefined batches, e.g. line of 20 continuous otolith in the box. Random selection of individual otoliths would be prone to errors.

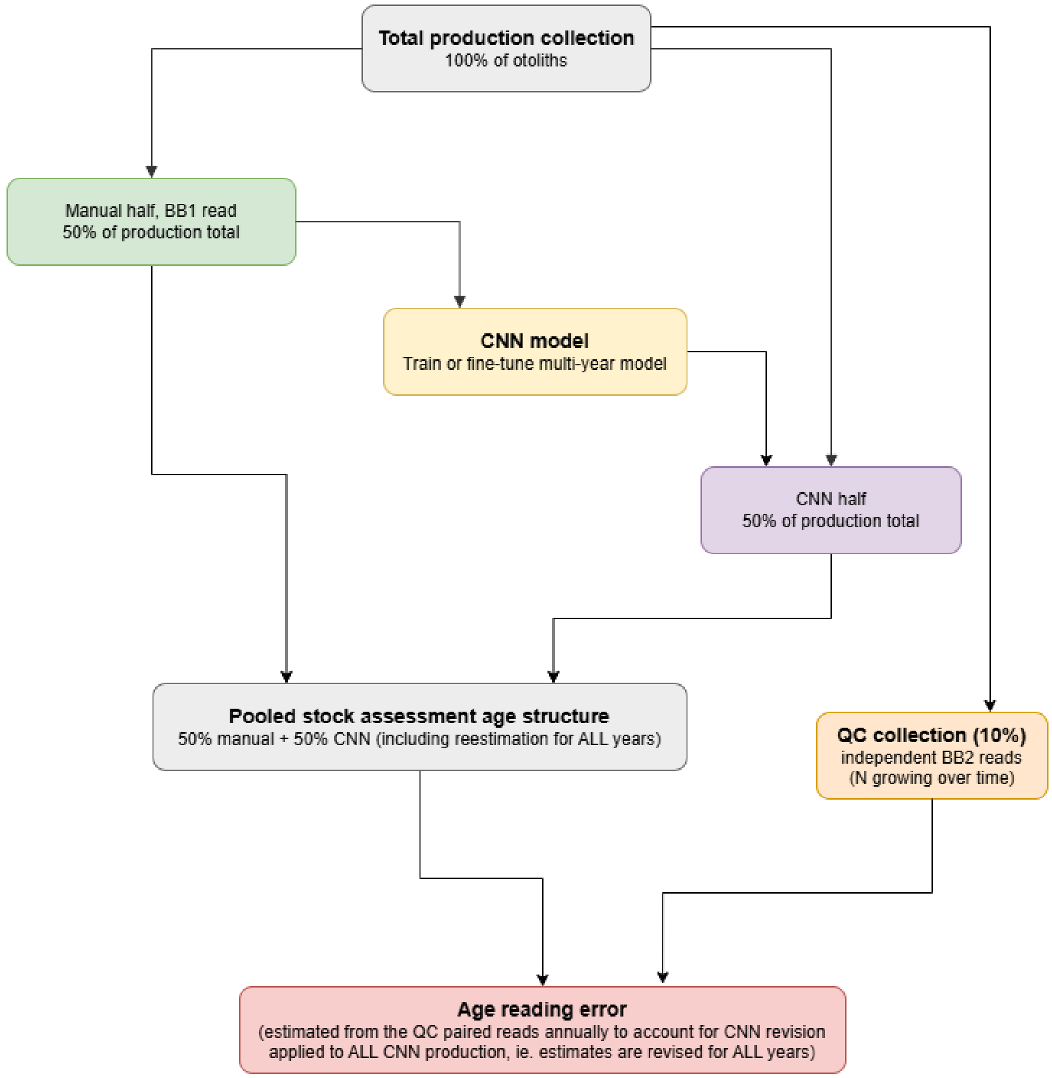

Figure 1: Mixed method approach flow chart

Ageing error of a 50/50 BB + AI mix — method test

A separate test was conducted to evaluate whether ages produced by a 50/50 random mix of break-and-bake (BB) reads and AI (CNN) predictions carry ageing error acceptable for use in the stock assessment. Ageing error was estimated following the method of Punt et al. (2008), as implemented in the AgeingError R package.[2] This approach fits paired age readings of the same otoliths to jointly estimate bias (systematic over- or under-ageing) and imprecision, with one reading method assumed to be unbiased. The output is an expected read age and associated standard deviation for each method at each true age, which can be summarized as a method-specific ageing-error matrix for use in the assessment.

For this analysis, BB was treated as the unbiased (based on the validation of Piner & Wischniowski, 2004) and the bias was estimated for both the mix method and previously for the historical surface ageing method. Imprecision was estimated for all three. The data included replicate reads for each combination of BB vs. BB, BB vs. surface and BB vs mix. All BB vs. mix method replicates were excluded from CNN model training (ConvNeXtSmall, 480, 3 seeds); the CNN was trained on 50% of the remaining images, simulating the 50/50 production split, and

[2] Version 2.2.0; https://pfmc-assessments.github.io/AgeingError/

used to predict ages for the held-out pairs. A dataset pairing BB with the 50/50 random draw of BB or the AI prediction was then used as input to the ageing-error estimation (Punt et al., 2008).

Figure 2 shows the estimated expected read age as a function of true age, with 95% confidence intervals, for re-estimated BB (black), surface ageing (red), and the 50/50 mix (blue), relative to the 1:1 (no-bias) line.

The results indicate that:

- The 50/50 mix has properties very close to BB-only ageing — slightly less precise and slightly biased, but substantially closer to BB than to the surface method.
- The BB vs. mix estimate is based on a relatively small replicate sample size (n = 1,587). The estimated bias-imprecision relationship may change, improving or worsening, with additional data spanning more years, and may also improve as AI models are further developed.

On the basis of currently available data, implementing a 50/50 random BB + AI mix appears both accurate and sufficiently precise for assessment use. However, further evaluation of temporal generalization — confirming that model predictions do not degrade over time — is needed before operational adoption.

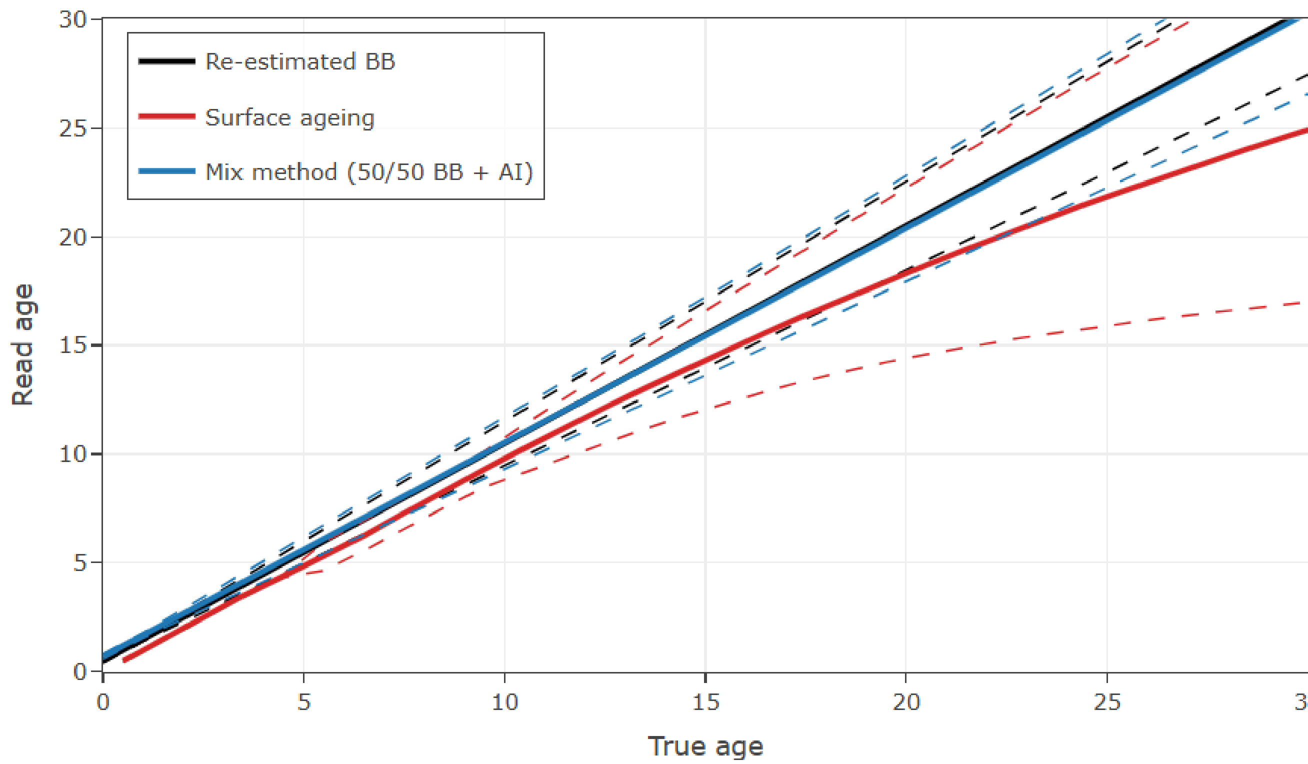

Figure 2: Expected read age by ageing method

Based on these findings, the IPHC plans as a next step to focus on supplementing the training and testing set with multi-year data spanning multiple decades. The IPHC's archive of previously aged otoliths provides a substantial resource for assembling such a set. Broadening the temporal range exposes the model to the inter-annual and inter-cohort variation arising from changing growth conditions and environmental regimes, which is precisely the source of the temporal generalization concern noted above.

Assembling and imaging this multi-decadal set is therefore a priority for demonstrating that model performance does not degrade over time and for supporting operational adoption of the mixed-method approach.

Evaluating new model architectures

Development to date focused mainly on InceptionV3 (the SRB028 model, Run 22). With the expanded 2019 image set, alternative backbones were re-evaluated on the held-out test set — ConvNeXtSmall, ConvNeXtBase, and EfficientNetV2M at 384 and 480 px, each averaged over three seeds. All of the newer architectures outperformed InceptionV3 across every metric, updating the earlier assessment (SRB026) that they underperformed — a difference attributable to the now considerably larger training set. The strongest configuration to date is ConvNeXtSmall at 480 px (Run 30). Results for individual model runs are available in Table 1.

Table 1: Comparison of model runs.

Run ID	22	26	28	**30**	32	34
Model	InceptionV3	ConvNeXtSmall	EfficientNetV2M	**ConvNeXtSmall**	ConvNeXtBase	ConvNeXtBase
Image size	400	384	480	**480**	384	480
Notes	SRB028			**Current favorite**		*In progress*
val_MSE	0.00116	0.00112	0.00110	0.00108	0.00111	
last epoch	192	191	122	142	169	
time [h] per run	15.6	35.2	18.7	38.7	50.3	
R^2	0.863	0.866	**0.871**	**0.871**	0.870	
RMSE	1.693	1.665	1.655	**1.639**	1.655	
MAE	1.081	1.025	1.034	**1.021**	1.054	
correctly predicted	33.5	**36.4**	35.4	35.7	33.5	
with +/- 1 tolerance	76.1	78.4	78.4	**79.3**	78.1	
within reader CI	69.1	72.1	72.1	**73.5**	70.6	
CV	6.07	5.68	5.80	**5.63**	5.93	
seeds	3 (3xrun time)	3 (3xrun time)	3 (3xrun time)	3 (3xrun time)	3 (3xrun time)	3 (3xrun time)

Detailed evaluation and performance metrics for the selected best-performing model specification are presented in Current results section and Appendix B: Deep ensemble individual results.

Assessing performance of using Z-stack images

Otolith images used for age prediction to date consisted of a single photograph taken at a focal plane selected manually at the time of imaging. Because the broken surface of a baked otolith is not perfectly flat, a single focal plane cannot hold the entire surface in focus, and part of the structure is necessarily rendered out of focus. The plane selected would also be expected to vary between images, introducing an uncontrolled source of variation. A Z-stack imaging capability was therefore developed, photographing each otolith at a sequence of focal depths rather than at a single plane.

Imaging was automated through a purpose-built application controlling a motorized linear translation stage coupled to the existing digital microscope camera. For each specimen the stage stepped through 80 focal positions at 0.1 mm intervals, recording one image at each position together with the specimen identifier, step number, and stage position. Camera settings and the stepping sequence were held constant across specimens so that the depth axis was comparable between otoliths.

Each completed stack passed an automated quality-control step. A focus measure (variance of the Laplacian) was computed for every slice to identify the sharpest slice and confirm that the sweep spanned the full range of focal depths containing the otolith surface. Stacks were then aligned by matching every slice to the middle slice of the stack, with the match estimated by maximizing the enhanced correlation coefficient (Evangelidis & Psarakis, 2008). This removed any lateral drift of the image across the sweep, arising if the stage travel axis was not perfectly parallel to the optical axis. A fixed window of 19 slices centered on the sharpest slice was then extracted for model input, giving every specimen the same depth extent positioned consistently relative to its own focal peak.

The modelling approach extended the existing two-dimensional architecture rather than replacing it. A pretrained ConvNeXtSmall backbone was applied independently to each slice, and the resulting per-depth features were combined by a learned attention mechanism over the depth axis and joined with the same biological and spatial covariates used in the two-dimensional models. The backbone was initialized from the corresponding single-image model, so that the comparison reflected the addition of depth information rather than a different starting representation. Attention weights were retained for each test specimen, indicating whether the model drew on multiple focal planes or concentrated on a single slice.

Performance was evaluated using a paired design in which four model variants across 3 inputs were trained and tested on identical specimens, identical training and validation folds, and identical random seeds, so that only the image input differed: (i) the full Z-stack; (ii) the single sharpest slice from that same stack, isolating the contribution of depth while holding specimen and camera constant; (iii) the conventional single image, retrained on the same specimens; and (iv) the deployed single-image model applied without retraining, representing current production performance. Of 1,057 specimens with both a Z-stack and a 2D image, 905 were used for training and validation and 152 were held out for testing. Specimens previously used to train the deployed model were placed in the training partition, leaving the test set unseen by all variants. The primary endpoint was the percentage of predictions falling within the empirical 95% confidence band of repeat human reads at the corresponding true age, with RMSE, MAE, R^2, exact agreement, and within-one-year agreement reported as secondary metrics; differences between variants were summarized as paired differences with bootstrap confidence intervals computed over the same test specimens.

This evaluation was a pilot conducted at substantially smaller scale than the primary model runs - 1,057 paired specimens compared with 11,107 images - and its conclusions therefore apply at that sample size. Z-stack imaging currently takes longer per specimen than photographing a single plane. However, the process is already fully automated once a specimen is positioned, and could be extended to unattended batch imaging of multiple otoliths, which would substantially reduce the effort per specimen and could ultimately prove more efficient than manually focusing each image individually. The purpose of the experiment was to determine whether the additional depth information yields a measurable improvement in age prediction sufficient to justify developing that capability.

Five versions of the model were tested on the same 152 otoliths that had been set aside for testing. They differed in how the model combined the images taken at different focal depths, in how many depths were used (19 or 9), in whether the underlying network was allowed to adjust during training, and, in the last version, in using conventional focus stacking to merge the whole set into one sharp image before the model saw it. None of them predicted age better than a model trained on the single sharpest image from the same stack. The differences were +0.0, −0.7, +0.7, −2.0 and +2.0 percentage points, averaging zero, and none could be distinguished from no difference at all. The focus-stacked image performed exactly the same as a single sharp image, even though it was measurably sharper. Repeating the same set-up produced differences of up to 3 percentage points from random variation in training alone, so none of the results above can be separated from that variation. The change in focus with depth is real, and in 88% of stacks the sharp region moved across the otolith as the focal plane advanced, but the model gained nothing from it beyond what the single sharpest image already provided. On this evidence, and at this sample size, capturing a Z-stack does not improve age prediction over a single well-focused image.

Z-stack imaging does, however, remove the need for a person to choose a focal plane. The manually focused images used to date performed at least as well as the sharpest image selected automatically from a stack, so manual focusing has not been limiting accuracy; the value of stack capture lies instead in making focus selection automatic and repeatable, and in opening the way to imaging otoliths in unattended batches. Whether the additional depth information becomes useful at higher input resolutions than the 384 pixels used here, at which much of the detail of a 2548-pixel image is discarded, has not been tested.

BACKGROUND

Otoliths are crystalline calcium carbonate structures, mostly in the form of aragonite, found in the inner ear of fish. They contain growth rings, that are often compared to tree growth rings. By analyzing the growth patterns in otoliths, scientists estimate the age of fish (Campana, 1999; Campana & Neilson, 1985), supporting the estimation of fish population demographics and population dynamics (Campana & Thorrold, 2001). In turn, fish age is a key input to stock assessment models that inform management decisions related to fish exploitation (Methot & Wetzel, 2013). It is estimated that the number of otoliths from captured fish that are read annually worldwide is on the order of one million (Campana & Thorrold, 2001).

The current method for determining ages of most fish species relies on manually extracting, preparing (embedding, sectioning), and reading otoliths. The simplest approach to reading the otolith is to immerse it in a clear liquid, such as water or alcohol solution, illuminate it from above, and view it against a dark background, using a stereo microscope. This method is suitable only for otoliths that are relatively thin with all annual bands visible from the surface. For species such as Pacific halibut, as the growth rate of the fish slows down, the outer growth bands become increasingly compressed and difficult to read from the surface of the whole otolith. To correctly determine the number of annual bands in such cases, otoliths are typically viewed in cross section which allows viewing the bands that are not visible from the surface view. In addition, the contrast between the growth rings can be enhanced through the baking process. Pacific halibut otoliths are aged using the ‘break and bake’ technique.

This manual ageing process is expensive, time-consuming,[3] and can be subject to bias[4] as well as imprecision due to variations in age estimations between readers and within readers over

[3] While the actual reading may account only for a fraction of the total cost and time required to process the otolith from collection to age determination, skilled readers require years of training, which should be considered when conducting a cost-benefit analysis.

[4] While the count of annual rings on Pacific halibut otoliths was found to provide unbiased age estimate using validation against bomb radiocarbon isotopes (Piner & Wischniowski, 2004), an earlier oxytetracycline (OTC) mark-recapture study indicated biases among age readers (Blood, 2003). In the 1980s, the IPHC applied injections with

time. Recent advances in imaging technologies and machine learning suggest that AI can assist in this process by automating the analysis of otolith images[5] and identifying and measuring the growth rings to determine age. AI algorithms can be trained on a large dataset of otolith images with known ages to learn the patterns and variations in growth rings. Once trained, the AI model can analyze new otolith images and predict the age of the fish based on the identified patterns in the image.

Using AI for age determination of Pacific halibut could improve consistency and replicability of age estimates, as well as provide time and cost savings to the organization, providing age data for reliable management advice. However, it's important to note that the AI model's accuracy depends on the quality and diversity of the training data, as well as the expertise of the scientists involved in training and validating the model. Regular validation and calibration with manual age determinations may be necessary to ensure the accuracy and reliability of the AI predictions. Thus, the proposed approach explores integrating AI-based age determination and traditional ageing methods for maximum accuracy of the estimates.

MODEL

Modeling approach

Previous literature (see perspective piece by Malde et al., 2020) suggests adapting a pre-trained convolutional neural network (CNN) designed for image classification to estimate age using otolith images obtained via microscope camera. This type of model is trained on a large collection of images of otoliths previously aged by human readers. Moen et al. (2018) presents the first case of the use of deep learning and CNN to estimate age from images of whole otoliths of Greenland halibut (*Reinhardtius hippoglossoides*).[6]

Artificial neural networks (ANNs) are computational structures inspired by biological neural networks. They consist of simple computational units referred to as neurons, organized in layers. The neuron parameters (or weights) are estimated by training the model using supervised learning. This process consists of two steps: forward propagation, where the network makes a prediction based on the input; and back propagation, where the network learns from its mistake by calculating the gradient of a loss function, and then uses the gradient to update the neuron weights. The ANNs approach has been used for fish ageing by Robertson & Morison (1999) and Fablet & Le Josse (2005) with a limited success.

The neural networks approach significantly improved in recent years with the increase in the number of layers, applying an approach often referred to as deep learning. Deep learning neural networks are known for their generality. With sufficient training data, they can be used to classify raw data (e.g., an array of pixels) directly, without explicit design of low-level features. The deep learning algorithm lower layers learn to distinguish between primitive features automatically, typically identifying sharp edges or color transitions. Subsequent layers then learn to recognize more abstract features as combinations of lower layer features, and finally merge this information to provide a high-level classification.

---

the antibiotic oxytetracycline (OTC) during routine tagging operations to evaluate validity of ageing method (IPHC, 1985). Upon injection, the OTC is absorbed by the fish's bony structure, including the otoliths, and leaves a mark that is easily seen when viewed under an ultraviolet light. When an OTC-injected tagged fish is recovered, the otoliths are removed and examined under the ultraviolet light. By comparing the number of annuli laid since the OTC mark to the fish recovery, the accuracy of the age readings can be determined.

[5] Although the idea of taking pictures of Pacific halibut otoliths is not new. See 1960 report by G. Morris Southward, *Photographing Halibut Otoliths for Measuring Growth Zones* (Southward, 1962).

[6] CNN was also applied for other tasks related to fisheries management, e.g. fish species identification (Allken et al., 2019).

In CNNs (LeCun et al., 1998; Simonyan & Zisserman, 2015), the layers are structured as stacks of filters, each recognizing increasingly abstract features in the data. Convolutional layers may be understood as an efficient way to transform an input image into another image, highlighting meaningful patterns, learned from data during training. The training is sequential, meaning the output of each layer is the input of the next layer, and the useful features are learned in the various layers during training. This approach is very effective for many image analysis problems, where objects are often recognized independent of their location. During network training, the performance is monitored over sequential epochs. Epochs represent the number of times that the training dataset is passed forward and backward through the network to refine model weights. Whenever the validation loss decreases, the trained model is saved, ending up with the network that corresponds to the minimum loss and highest accuracy on the validation set. The trained network is then evaluated on the testing set.

In the CNN model, age prediction from otolith images can be formulated either as a classification task - where age is treated as a categorical variable - or as an image regression task, which involves predicting a continuous numerical value. Although treating fish age as a discrete parameter is a common method for identifying individual year classes, i.e., grouping fish by spawning year (Moen et al., 2018), this approach has proven less effective for Pacific halibut. As a long-lived species with a wide distribution of age classes, Pacific halibut pose a challenge for classification-based methods. The oldest Pacific halibut on record have been aged at 55 years (Keith et al., 2014).

Software and model architecture options

The proposed approach builds on prior work by Moen et al., (2018) and Moore et al., (2019), who implemented CNNs for otolith-based fish age estimation using the TensorFlow and Keras libraries. TensorFlow remains one of the most widely used and well-supported frameworks for deep learning, and Keras provides a high-level API that simplifies model development.

The approach uses transfer learning — repurposing a model pre-trained on a large, related task rather than training from random initialization, which substantially improves performance when labelled data are limited. Development initially adapted the InceptionV3 model (pre-trained on the ImageNet database of over 14 million annotated images) following the open-source implementation of Politikos et al. (2022) for Greek red mullet (*Mullus barbatus*) ageing (github.com/dimpolitik/DeepOtolith), with the classification output replaced by a single numeric output for regression.

To adapt a pre-trained backbone for Pacific halibut ageing, the input layer is scaled to the otolith image resolution[7] and the output is changed from class probabilities to a single continuous value.[8] The general pattern is: Input → pre-trained backbone (feature extractor) → regressor → age output, with capture covariates integrated through a parallel branch and fused before the

[7] Image resolution refers to an image's pixel dimensions (width × height in pixels). The original otolith images (2,548 × 2,548 pixels) were downsized to the model's input resolution.

[8] Alternatively, Politikos et al. (2021) replaced the last layer with a feed-forward network with two hidden layers replacing the default 1000-categories output layer with a fully-connected layer with six hidden nodes, corresponding to a limited number of age categories [Age-0 – Age-5+], with the last one representing fish of age 5 and older. In this case, the network outputs probabilities using the softmax function, a function that performs multi-class classification and transforms the outputs to represent the probability distributions over a list of potential outcomes. The IPHC uses in its stock assessment bins Age-2 – Age 25+ for the current age data and Age-2 - Age-20+ for the historical surface read ages. The adoption of a larger number of age categories prompted the decision to incorporate a regression layer in place of class probabilities.

final prediction (see Use of covariates). Models are trained with the Adam optimizer to minimize mean squared error (MSE) between predictions and expert annotations.

A range of backbone architectures has been evaluated to identify the best trade-off between predictive performance and computational cost: InceptionV3, EfficientNet variants (EfficientNetB4, B5, and EfficientNetV2 M/L), and ConvNeXt (Small and Base). EfficientNet architectures balance the scaling of depth, width, and resolution for efficiency and accuracy, with EfficientNetV2 adding progressive training and improved scaling. ConvNeXt architectures modernize convolutional design with elements inspired by transformer models, achieving competitive accuracy with a comparatively simple structure.

An earlier assessment (IPHC-2025-SRB026-10) found these alternative architectures underperformed relative to InceptionV3, with several EfficientNet and ConvNeXt configurations exhibiting limited learning and predictions regressing toward the mean age - behavior consistent with insufficient training data for the larger models at that time. With the subsequent expansion of the image database, this conclusion no longer holds: on the current dataset, ConvNeXt and EfficientNetV2M match or exceed InceptionV3 across all evaluation metrics.

The current preferred configuration is **ConvNeXtSmall at 480×480 resolution**, which gives the best overall results to date (lowest RMSE and highest agreement with expert ages), narrowly ahead of EfficientNetV2M at 480 - the latter reaching comparable accuracy at roughly half the training time. Increasing model capacity further (ConvNeXtBase) did not improve accuracy while increasing compute cost, indicating that dataset size, rather than model size, is the current limiting factor. Architecture evaluation remains ongoing as the dataset grows.

TensorFlow/Keras remains the primary framework for the current implementation. Future work may also explore alternative frameworks such as PyTorch (originally developed by Meta), which offers flexible dynamic computation graphs and growing adoption in the research community.

Performance metrics and achieved accuracy

Model performance on the test set is assessed by comparing the rounded regression output against expert ages using root mean squared error (RMSE), mean absolute error (MAE), the percentage of exact matches, and the percentage within ±1 year. Following Moen et al., (2018), the coefficient of variation (CV) is also reported, allowing comparison against the imprecision of manual reading.[9]

Published CNN ageing studies provide useful benchmarks. Moen et al., (2018), for Greenland halibut, aged 29% of otolith pairs exactly and 38% within one year (n=164), and noted a tendency to under-age older fish with prediction variance increasing with age. Moore et al. (2019), using CT scans,[10] matched the human reader exactly for 47% of snapper, with a further 35% of ages estimated within 1 year of the human reader estimate of age (n=687). For hoki, the model gave the same age as the human reader for 41% of individuals (n=882). The age model for Greenland halibut by Politikos et al., (2022) gave RMSE of 1.69 years between age prediction and age

[9] The CV of the predicted age at true age is the primary input to the IPHC stock assessment. It is generally modelled as a parametric function of age accounting for the complex joint probability that both estimates can be incorrect (Punt et al., 2008).

[10] CT scanning uses X-ray technology to produce image slices through objects, which can be reconstructed into virtual, three-dimensional (3D) images that can be rotated and viewed in any orientation (Moore et al., 2019). Such images may provide more accurate estimates, but the cost of this approach is prohibitive at (based on trial conducted in New Zealand) $1,500 per day, with scan timed for an individual otolith between 40 min to one hour. However, as the technology progresses, this approach may provide an option for fully automating the entire ageing process by scanning a whole fish (e.g., along a conveyor belt). Deep learning methods (i.e., CNN) developed for age determination from surface images could serve as a base for age determination from CT scans.

reading by experts (n=8,218, 26 age categories). For Greek red mullet, correct age was predicted for 69.2% individuals, with an additional 28.2% being within 1 year of error (n=5,027). Benson et al., (2023), using near-infrared spectroscopy with geospatial and biological covariates for walleye pollock, achieved a test-set RMSE of 0.91, with at least 67% of ages within ±1 year.

These figures should be read against the imprecision of manual ageing itself, which is not perfectly accurate. For Pacific halibut, within- and between-reader agreement is generally 60–70% exact, 80–90% within one year, and 100% within three years (see *Technical Report No. 46* and *No. 47*). This sets a practical ceiling: a model that agrees with expert ages about as often as two readers agree with each other is effectively performing at the level of the method it supplements.

Database

The IPHC annually ages a considerable number of otoliths (see *Appendix A* for details). Since 1925, over 1.6 million otoliths have been aged and stored for potential future use. Otoliths collected by the IPHC for ageing purposes undergo additional processing. Otoliths are sectioned (broken in half) and baked to enhance the contrast between the growth rings. These stored and previously aged otoliths serve as a valuable resource for creating a database of images for training purposes. To optimize model training, the selection of otoliths included in the model covers a broad spectrum of fish sizes, ages, sexes, and collection locations.

Before photographing, processed otoliths were placed in a monochrome tray featuring an elongated groove designed to keep the otolith upright and immersed in water. The pictures were taken with AmScope 8.5MP eyepiece cameras,[11] under consistent lighting conditions and magnification. The input database includes images of standardized size, 2,548 by 2,548 pixels, which are later resized to the desired resolution based on the model's specification.[12]

It is important to note that <u>it may not be necessary to image the otoliths at resolutions sufficient for human viewers to resolve, because the CNN may be able to arrive at an age estimate without directly counting bands</u> (Moore et al., 2019).

Figure *3* shows an example of a range of images used in the CNN training dataset.

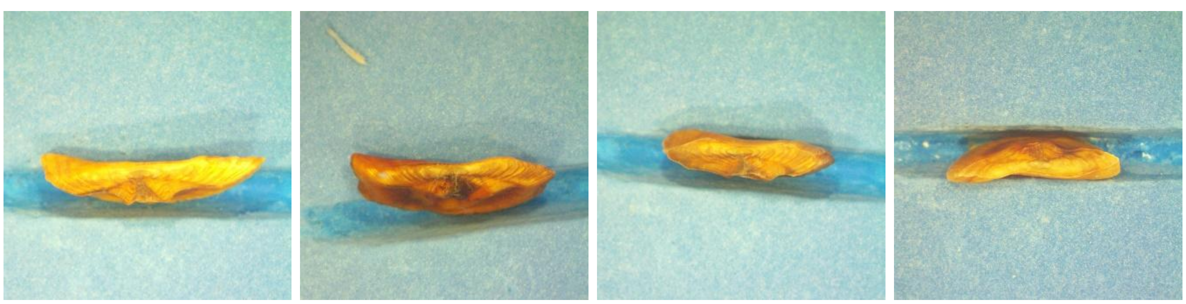

Figure 3. Examples of Pacific halibut otolith images taken for inclusion in the training set.

To assess whether the break-and-bake (BB) processing step is necessary for AI-based ageing, a complementary set of otolith images captured prior to processing (surface images) was assembled and used to train models directly comparable to the BB models. This supports a cost-benefit comparison, since surface imaging avoids the time consuming sectioning and baking. Example images are provided in Figure *4*. The approach mirrors Politikos et al. (2022), who used

[11] The camera fits in one of the microscope eyepieces, eliminating the need to purchase a separate camera mount for the microscope.

[12] Higher resolution images offer the flexibility to adapt the model in the future to more detailed and complex image analysis tasks, potentially improving the accuracy and effectiveness of image recognition capabilities.

unprocessed otoliths imaged in water under a stereomicroscope on a white background with transmitted light.

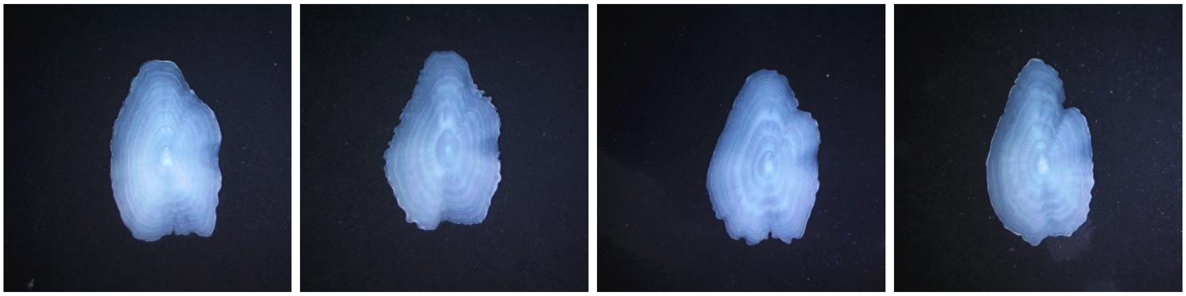

Figure 4. Examples of Pacific halibut otolith images taken for inclusion in the training set.

Presorting otoliths

The adopted procedure excludes broken otoliths, applying manual presorting at the image-taking stage. Presorting has also occurred at the collection stage when crystalized otoliths are omitted when collecting samples.[13]

Image collection

The image collection is associated with labels storing:

1. Otolith reference number – using referencing system already in place;
2. Image name and location – exact path for image access;
3. Resolved age – human reader derived age (**rsvage**);
4. Label – number of individual reads – serves as a measure of ageing quality.
5. Year collected – to account for variation between cohorts and prevalent environmental conditions;[14]
6. Date collected – to account for the 'edge effect' reflecting seasonal changes;
7. Geospatial characteristics of the collection site (latitude, longitude and IPHC Regulatory Area) – to capture regional variation;
8. Resolved sex – to determine whether otolith characteristics (possibly not directly visible to human eye) could be used for sex determination.[15]

Uncertainty estimates

Several approaches were considered for quantifying model uncertainty, including Monte Carlo dropout (Gal & Ghahramani, 2016) and deep ensemble method (Lakshminarayanan et al., 2017). Following evaluation, **deep ensemble method** was identified as the more suitable approach for the present application. Deep ensembles involve training multiple independently initialized models and aggregating their predictions to produce a consensus estimate, with prediction variance across ensemble members serving as a measure of uncertainty. Compared with Monte Carlo dropout, deep ensembles generally provide more stable predictive performance, improved calibration of confidence estimates, and greater robustness to out-of-distribution or ambiguous samples. Although computationally more demanding, this approach

[13] Crystalized otoliths have an altered composition – specifically, where the aragonite in the otolith is partially or mostly replaced by vaterite, a phenomenon known as otolith crystallization. Crystallized otoliths are not suitable for ageing. About 1% of otoliths are partly crystallized and are assigned ages. The same is true for broken otoliths that are aged (1%)

[14] Year collected is currently not incorporated as a model covariate due to the limited temporal range represented within the training dataset. Future model versions may re-evaluate its inclusion as additional years of data become available.

[15] IPHC is currently using genotyping for Pacific halibut sex determination.

better aligns with Pacific halibut ageing workflows, where reliable identification of uncertain predictions in a production setting will be essential for the development of a semi-automated, quality-controlled ageing protocol that leverages the strengths of both AI and human expertise.

CURRENT RESULTS

Selected model evaluation

The selected model configuration (ConvNeXtSmall) utilized 11,107 images of otoliths collected during the 2019 IPHC fishery-independent setline survey (FISS). The 2019 FISS represents a comprehensive sampling effort expected to reflect regional variability in Pacific halibut otolith characteristics. As such, it provides a robust foundation for initial model development and evaluation.

The images were divided into training, validation, and test datasets. The training set (7,740) was used for training purposes. The validation set (1,367) was used to evaluate the model during the training process, allowing for adjustments without using the test set, which was reserved for the final evaluation. The test dataset (2,000) was used to assess the performance of the model after training, providing an unbiased evaluation of its generalization capability to new, unseen data. All images were resized to 480x480 pixels. Images of broken otoliths were excluded.

The selected model employed a maximum of 600 training epochs, with early stopping patience set to 60 epochs. A learning rate reduction was triggered if validation loss plateaued for 30 epochs, reducing the rate by a factor of 0.8. The initial learning rate was set at 0.0002, and training was performed using a batch size of 8. A comprehensive suite of image augmentation techniques (e.g., rotation, position shift, zoom, brightness variation) was applied to improve generalization and robustness.

To enhance model reliability and quantify uncertainty, a deep ensemble approach was adopted. The model was trained 7 times, each with a different random seed. Ensemble outputs were averaged to produce final predictions and calculate prediction uncertainty. Detailed results for individual ensemble members are provided in Appendix B.

Across ensemble runs, the model trained for an average of 161 epochs (101 effective epochs with early stopping set at 60). It achieved a normalized MSE average of 0.001003 on the validation set. When averaged across seeds results were rounded to the nearest integer age, the RMSE for the test set was 1.621 and MAE was 0.998. On average, the ensemble predicted the exact age correctly for 36.7% of test images, and an additional 43.5% were within ±1 year of the manually assigned age, resulting in a total agreement within 1 year for 80.2% of cases. 74.4% of predictions were within manual ageing confidence interval.

Figure *5* shows a comparison between manually derived ages and AI-predicted ages across the ensemble. Figure 6 compares the age composition estimated manually with that derived from the ensemble model predictions.

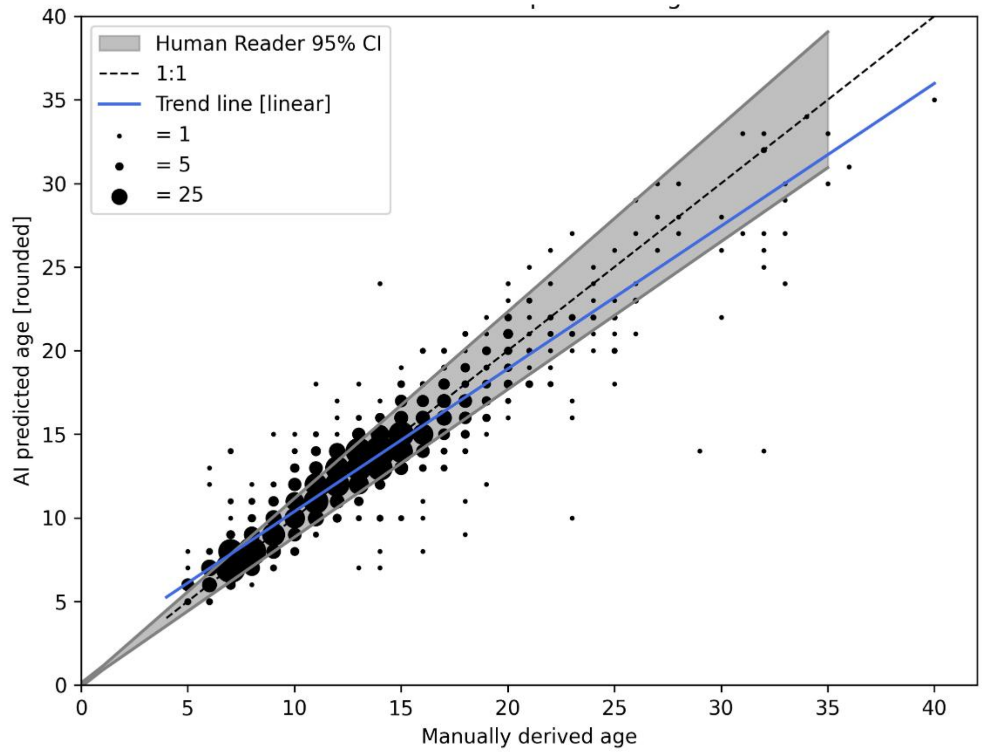

Figure 5. Comparison between manually derived age with AI predicted age.

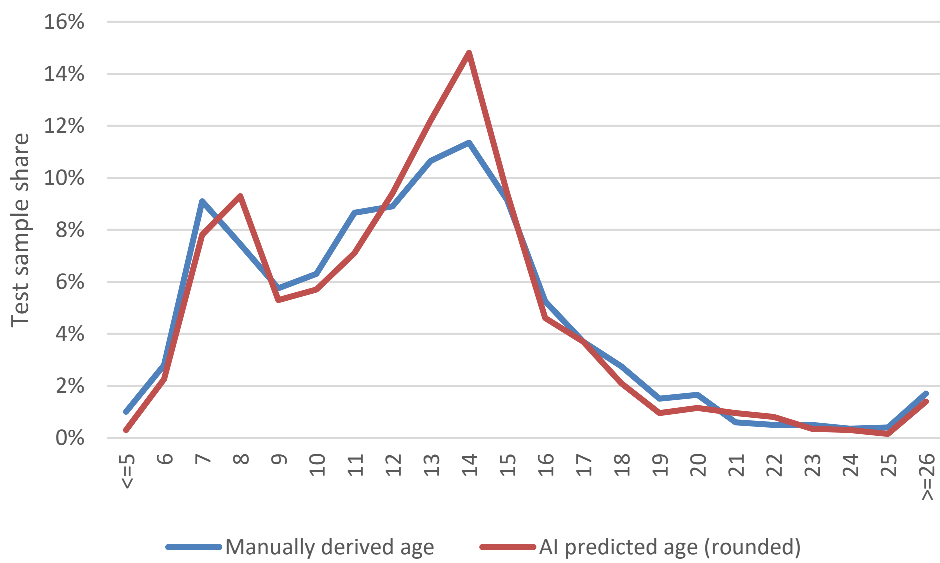

Figure 6. Comparison between manually derived age with AI predicted age – age composition.

It is important to note that although the model tends to underestimate the ages of older Pacific halibut on average, the statistically significant bias previously observed for age categories 21+ (IPHC-2025-SRB026-10) is no longer apparent. The number of observations for older age categories remains low despite an overall increase in sample size (Figure 7). This suggests that the saturation point for achieving optimal accuracy in older age categories may not yet have been reached, and the model could benefit from further improvement by adding more images representing older age categories to the training set. Currently, only 4.1% of the otoliths used in the model were from fish aged 21 or older.

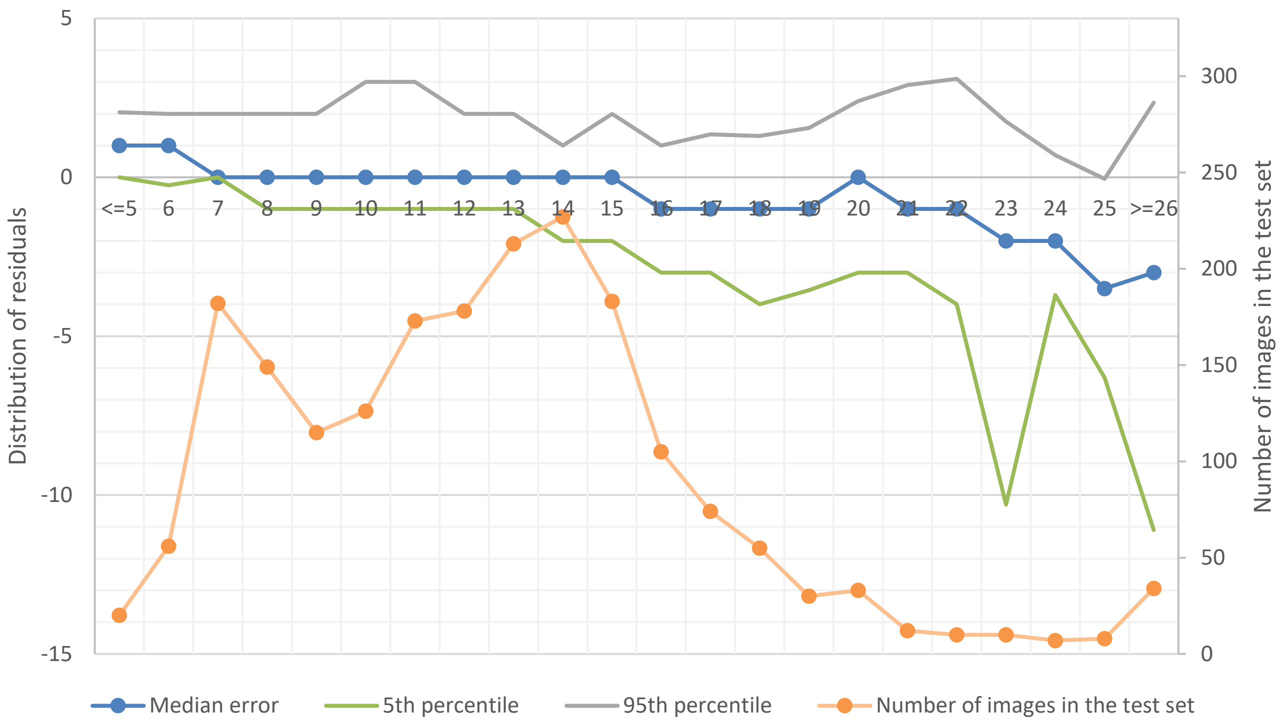

Figure 7. Distribution on residuals and number of images by age in the test set.

Temporal generalization

A key requirement for operational use is temporal generalization, that is, the ability of a model trained on one year's otoliths to age otoliths collected in later years, for which no training data are available. In earlier work, a model trained on the 2019 FISS sample showed some decline in accuracy when applied to a later survey year, relative to its performance on a test set from the same year. However, as this is based on a single pair of years, the observed decline could reflect inter-annual variability, changes in the underlying stock or sampling, or other factors specific to those years, rather than a systematic trend over time.

A key strength of the proposed design is that the 50/50 mix supplies a fresh set of manual (BB1) reads every year. This continually supplements the training set with labelled data from the current year, and CNN ages are recalculated when the model is updated, so predictive accuracy is expected to be maintained, and to improve, through the ongoing addition of data from the current year (and, in future, data spanning multiple decades) rather than through calibration to any single year.

The deep ensemble framework provides an additional safeguard. Uncertainty for each sample, measured as the standard deviation across ensemble members, allows predictions to be sorted by confidence, so that the most reliable ages are used directly and uncertain cases are flagged for expert review. Fine-tuning on a subset of otoliths from new years, optionally targeting the most uncertain samples, was previously shown to recover accuracy and could serve as one way of adjusting the model to new years. In the proposed design, however, the main route to robustness across years is the continual expansion of the training set rather than fine-tuning on its own.

No new temporal-generalization testing was undertaken in the current update, which focused on model architecture and the mixed-method design. Extending this assessment to the current preferred model and across additional years remains a priority, and is closely tied to the planned expansion of the training set with multi-year data.

Surface images use

Using the same architectures and covariates, models trained on surface images performed worse than the equivalent BB models, though the gap was moderate. The best surface configuration (ConvNeXtSmall at 480 px; 3 seeds) reached a test-set RMSE of 1.89 years with 76.0% of predictions within ±1 year, compared with RMSE 1.73 and 77.2% for the corresponding BB model; across configurations, surface RMSE was consistently about 0.15 years higher. Exact-match agreement and R^2 were comparable between the two image types, indicating that surface models predict the bulk of younger ages about as well but make larger errors on older otoliths - consistent with surface images being unable to resolve the compressed outer growth increments that the break-and-bake cross-section is specifically intended to reveal. These results indicate that, at present, surface imaging alone does not match the accuracy of processed otoliths, particularly for older fish.

CONCLUSIONS

The continued advancement of AI for otolith-based ageing offers considerable potential to improve the efficiency of Pacific halibut age determination. The results presented here indicate that CNNs, implemented as a deep ensemble, now achieve predictive accuracy that supports their use as a supplement to, and within a structured mixed-method design a partial replacement for, the current manual ageing protocol. On the 2019 test set, the selected seven-member ensemble (ConvNeXtSmall at 480x480 pixel image resolution) agreed with expert ages within one year for 80.2% of otoliths (36.7% exact) and fell within the human-reader confidence interval for 74.4%, approaching the level of agreement expected between human readers.

Continued expansion of the image database remains the single most important driver of progress. Increasing the training set from 2,799 to 11,107 otolith images improved representation across age classes and enabled the broader evaluation reported here. Refinements to the modelling pipeline contributed incrementally, and a systematic comparison of architectures showed that newer backbones (ConvNeXt and EfficientNetV2M) now outperform the InceptionV3 model used previously, with ConvNeXtSmall at 480 pixels selected as the current preferred configuration. Auxiliary covariates (capture location and collection date) again produced modest but consistent gains.

A central contribution of this update is the proposed 50/50 mixed-method design, in which half of each year's production is aged by a single expert break-and-bake read and half by the CNN, with the two pooled for the stock assessment. A formal ageing-error analysis showed that ages from the 50/50 mix carry error close to that of break-and-bake ageing, and appear acceptable for assessment use at the current data volume. The design preserves total ageing effort at current levels while roughly halving the specialist manual burden, builds in the quality-control second reads needed to monitor ageing error over time, and supplies a fresh set of expert reads each year that continually supplements the training set with current-year data.

This annual supply of expert reads also makes the 50/50 design a mechanism for adapting to imperfect temporal generalization. The model is updated with current-year labelled data every year, so any inter-annual change or emerging trend is captured directly in the training data and carried through the updated predictions, with CNN ages re-estimated for all years once the model is updated. Earlier work suggested some loss of accuracy when a model trained on one year was applied to another without such updating; this rests on a single pair of years and, in any case, describes the static-model situation the design avoids. No new temporal-generalization testing was undertaken in this update. Extending the training and test data across additional years, drawing on the IPHC's otolith archive, would further confirm performance across years and support progression toward fuller operational reliance.

Two further lines of work returned informative results. Models trained on unprocessed surface images performed worse than break-and-bake models, particularly for older otoliths, confirming that break-and-bake processing remains necessary. A pilot evaluation of Z-stack (multi-focal-depth) imaging showed no accuracy gain over a single well-focused image at the resolution tested, although it removes the need for manual focus selection and enables automated batch imaging.

Overall, the ageing-error assessment is the key result supporting operational use: ages from the 50/50 mix are slightly less precise and slightly biased relative to break-and-bake ageing, but substantially closer to break-and-bake than to the historical surface method, and acceptable for assessment use based on currently available data.

Finally, AI-based ageing will continue to depend on human expertise, both to generate the validated training data that capture temporal and environmental change and to provide ongoing oversight. Integrating expert reading with continual model updating within the mixed-method workflow is central to reliable operational use.

AI USE DISCLOSURE

The concept, design, and execution of this work, were carried out by the author. Generative AI tools provided support with coding, background research, and text review and refinement. The author reviewed all AI-assisted output and is responsible for the final content.

ACKNOWLEDGEMENTS

The author thanks the interns and volunteers who assisted the IPHC with photographing otoliths, providing the images used as input to the modelling described here.

LITERATURE

Allken, V., Handegard, N. O., Rosen, S., Schreyeck, T., Mahiout, T., & Malde, K. (2019). Fish species identification using a convolutional neural network trained on synthetic data. *ICES Journal of Marine Science*, *76*(1), 342–349. https://doi.org/10.1093/icesjms/fsy147

Benson, I. M., Helser, T. E., Marchetti, G., & Barnett, B. K. (2023). The future of fish age estimation : deep machine learning coupled with Fourier transform near-infrared spectroscopy of otoliths. *Canadian Journal of Fisheries and Aquatic Sciences*, *00*, 1–13. https://doi.org/dx.doi.org/10.1139/cjfas-2023-0045

Blood, C. L. (2003). I . Age validation of Pacific halibut II . Comparison of surface and break-and-burn otolith methods of ageing Pacific halibut. *IPHC Technical Report*, *47*.

Campana, S. E. (1999). Chemistry and composition of fish otoliths: Pathways, mechanisms and applications. *Marine Ecology Progress Series*, *188*, 263–297. https://doi.org/10.3354/meps188263

Campana, S. E., & Neilson, J. D. (1985). Microstructure of Fish Otoliths. *Canadian Journal of Fisheries and Aquatic Sciences*, *42*(5), 1014–1032. https://doi.org/10.1139/f85-127

Campana, S. E., & Thorrold, S. R. (2001). Otoliths, increments, and elements: keys to a comprehensive understanding of fish populations? *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(1), 30–38. https://doi.org/10.1139/f00-177

Evangelidis, G. D., & Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(10), 1858–1865. https://doi.org/10.1109/TPAMI.2008.113

Fablet, R., & Le Josse, N. (2005). Automated fish age estimation from otolith images using statistical learning. *Fisheries Research*, *72*(2–3), 279–290. https://doi.org/10.1016/j.fishres.2004.10.008

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*.

IPHC. (1985). Annual Report 1984. In *IPHC Annual Report*.

Keith, S., Kong, T., Sadorus, L. L., Stewart, I. J., & Williams, G. (2014). The Pacific halibut: biology, fishery, and management. *IPHC Technical Report*, *59*. https://doi.org/10.1042/bj0490062

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient Based Learning Applied to Document Recognition. *Proc. of the IEEE*.

Malde, K., Handegard, N. O., Eikvil, L., & Salberg, A. B. (2020). Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, *77*(4), 1274–1285. https://doi.org/10.1093/icesjms/fsz057

Methot, R. D., & Wetzel, C. R. (2013). Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research*, *142*, 86–99. https://doi.org/https://doi.org/10.1016/j.fishres.2012.10.012

Moen, E., Handegard, N. O., Allken, V., Albert, O. T., Harbitz, A., & Malde, K. (2018). Automatic interpretation of otoliths using deep learning. *PLoS ONE*, *13*(12), e0204713.

Moore, B. R., Maclaren, J., Peat, C., Anjomrouz, M., Horn, P. L., & Hoyle, S. (2019). Feasibility of automating otolith ageing using CT scanning and machine learning. *New Zealand Fisheries Assessment Report*, *58*.

Ordoñez, A., Eikvil, L., Salberg, A. B., Harbitz, A., Murray, S. M., & Kampffmeyer, M. C. (2020). Explaining decisions of deep neural networks used for fish age prediction. *PLoS ONE*, *15*(6), 1–19. https://doi.org/10.1371/journal.pone.0235013

Piner, K. R., & Wischniowski, S. G. (2004). Pacific halibut chronology of bomb radiocarbon in otoliths from 1944 to 1981 and a validation of ageing methods. *Journal of Fish Biology*, *64*(4), 1060–1071. https://doi.org/10.1111/j.1095-8649.2004.0371.x

Politikos, D. V, Petasis, G., Chatzispyrou, A., Mytilineou, C., & Anastasopoulou, A. (2021). Automating fish age estimation combining otolith images and deep learning: The role of multitask learning. *Fisheries Research*, *242*, 106033. https://doi.org/https://doi.org/10.1016/j.fishres.2021.106033

Politikos, D. V, Sykiniotis, N., Petasis, G., Dedousis, P., Ordoñez, A., Vabø, R., Anastasopoulou, A., Moen, E., Mytilineou, C., Salberg, A. B., Chatzispyrou, A., & Malde, K. (2022). DeepOtolith v1.0: An Open-Source AI Platform for Automating Fish Age Reading from Otolith or Scale Images. *Fishes*, *7*(3), 1–11. https://doi.org/10.3390/fishes7030121

Punt, A. E., Smith, D. C., KrusicGolub, K., & Robertson, S. (2008). Quantifying age-reading error for use in fisheries stock assessments, with application to species in Australia's southern and eastern scalefish and shark fishery. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(9), 1991–2005. https://doi.org/10.1139/F08-111

Robertson, S. G., & Morison, A. K. (1999). A trial of artificial neural networks for automatically estimating the age of fish. *Marine and Freshwater Research*, *50*(1), 73–82. https://doi.org/10.1071/MF98039

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR 2015 - Conference Track Proceedings*.

Southward, G. M. (1962). Photographing Halibut Otoliths for Measuring Growth Zones. *Journal of the Fisheries Research Board of Canada*, *19*(2), 335–338. https://doi.org/10.1139/f62-018

Yang, Y., Zha, K., Chen, Y. C., Wang, H., & Katabi, D. (2021). Delving into Deep Imbalanced Regression. *Proceedings of Machine Learning Research*, *139*, 11842–11851.

APPENDIX A: COUNTS OF OTOLITHS AGED BY THE IPHC

Collection year	Ageing method	IPHC FISS*	Commercial (Market Sample)*	NOAA Trawl survey*	Tag recovery*	ADF&G recreational*	Clean collection
pre-1960	surface	70,984			10,068		
1960	surface	6,606			681		
1961	surface	4,727		4,576	842		
1962	surface	2,605		1,692	594		
1963	surface	8,257		2,209	440		
1964	surface	10,295	27,828	1,001	353		
1965	surface	5,169	27,252	1,186	493		
1966	surface	3,750	24,638	1,777	796		
1967	surface	6,325	29,797	2,271	1,151		
1968	surface	2,314	29,772	1,887	1,813		
1969	surface	1,510	23,361	1,019	1,869		
1970	surface	1,138	24,686	1,184	867		
1971	surface	2,702	16,374	2,294	732		
1972	surface	2,597	23,381	1,180	490		
1973	surface	1,747	16,683	893	244		
1974	surface	1,021	11,569	1,189	128		
1975	surface	1,212	14,128	1,136	131		
1976	surface	1,843	14,103	969	72		
1977	surface	1,853	13,514	1,102	83		
1978	surface	1,933	11,434	1,309	61		
1979	surface	2,021	7,219	730	93		
1980	surface	5,022	10,317	717	168		
1981	surface	7,942	8,267	460	129		
1982	surface	5,720	9,644	443	208		
1983	surface	5,822	9,262	1,355	286		
1984	surface	6,508	10,233	1,089	455		
1985	surface	5,872	12,986	1,192	778		
1986	surface	5,139	12,426	1,120	1,020		
1987	surface	42	16,137		859		
1988	surface	1,179	17,154	98	761		
1989	surface	6,130	14,122		710		
1990	surface	2,201	14,800	4,802	397		
1991	surface	1,315	13,461	2,598	280		
1992	surface/BB	7,530	14,564	222	182		
1993	surface/BB	3,384	13,747		147		
1994	surface/BB	2,618	13,311		99		
1995	surface/BB	4,512	12,297	433			
1996	surface/BB	10,893	13,452	2,211			
1997	surface/BB	14,784	15,501	834	148		
1998	surface/BB	8,587	14,395	1,145	98		

1999	surface/BB	11,971	12,858	3,029	70	3,672	
2000	surface/BB	14,122	13,982	1,209	46	2,706	
2001	surface/BB	14,731	13,181	2,952	27	2,609	
2002	BB	13,635	17,932	761	24	2,349	
2003	BB	12,626	13,915	3,876	79	2,754	
2004	BB	14,474	11,798	897	450	3,288	
2005	BB	12,651	14,650	2,028	643	3,183	
2006	BB	14,976	13,399	2,621	679	3,179	
2007	BB	16,285	13,964	3,930	455	3,026	
2008	BB	15,545	13,460	1,527	304	1,500	
2009	BB	15,706	13,583	4,922	276	1,500	
2010	BB	14,080	16,106	1,915	21	1,500	625
2011	BB	14,451	11,391	4,592	26	1,500	676
2012	BB	17,896	12,902	1,639	9	1,500	1164
2013	BB	12,717	11,039	2,044	19	1,503	1020
2014	BB	16,194	12,606	1,476	22	1,500	1096
2015	BB	15,815	12,312	2,133	24	1,500	1072
2016	BB	15,113	11,618	742	21	1,502	902
2017	BB	12,565	10,821	1,384	15	1,500	756
2018	BB	12,935	11,013	576	39	1,499	798
2019	BB	17,716	10,711	1,640	34	1,497	925
2020	BB	10,323	10,568	-	34	1,413	577
2021	BB	12,253	11,051	1,444	38	1,500	547
2022	BB	9,702	10,942	1,902	39	2,334	519
2023	BB	8,506	10,932	(3,147)	(48)	(1,958)	462
2024	BB	5,770	10,474[1]	1,058	(61)	1,542[2]	458
2025	BB	7,912[3]	9,740[4]	(2,379)	(35)	(1,456)	499

Notes:

- Star (*) indicates blind side otolith.
- BB stands for ‘break and bake’ approach.
- All otoliths reported in this table were aged with the exception of the clean collection.
- All aged otoliths are stored in glycerol/thymol solution.
- Some small fish from trawl survey collection are still aged by surface method; otoliths with surface age>4 are sectioned and baked.
- Sample data not entered prior to 1960 for FISS, 1964 for commercial, 1961 for NOAA trawl survey.
- Clean collection is not aged, stored dry, and include paired otoliths.
- Tribal otoliths are included in the Market Sample series.
- Additionally, there are 144 not aged 2A recreational otoliths, all from Hein Bank collected between 2004 and 2009.
- Sex information available since 2017 (typically ca. 1 year of lag).
- Trawl and recreational otoliths lag one year in ageing.
- In brackets, otoliths available for ageing but ageing not completed.

[1] Commercial otolith collection subsampled: 10,474 otoliths were collected, 7,057 were selected for ageing.
[2] 2024 ADF&G recreational otolith collection subsampled: 1,542 otoliths were collected, 819 were selected for ageing.
[3] 7,912 FISS total for 2025 includes 242 fecundity samples. Some otoliths from Area 4A were subsampled: 7,670 were selected for ageing.
[4] 2025 commercial otolith collection subsampled: 9,740 otoliths were collected, 6,723 otoliths were selected for ageing.

APPENDIX B: DEEP ENSEMBLE INDIVIDUAL RESULTS

Model run	1	2	3	4	5	**6**	**7**	
Seed	11	27	42	44	51	**62**	**66**	**ALL SEEDS mean**
RESULTS – TEST SET								
Test RMSE	1.678	1.748	1.767	1.741	1.711	1.694	1.673	1.621
Test MAE	1.038	1.101	1.117	1.118	1.046	1.066	1.082	0.998
Test CV	5.680	6.097	6.178	6.146	5.748	5.907	6.032	5.494
Test R^2	0.865	0.853	0.850	0.855	0.856	0.864	0.865	0.876
Correctly predicted	36.0	33.9	32.7	32.7	36.5	34.9	32.6	36.7
Correctly predicted with ±1 year tolerance	78.6	76.6	76.4	75.1	78.4	76.7	76.4	80.2
Predicted within reader confidence interval	73.2	70.5	70.0	69.6	72.4	71.2	70.1	74.4

* Indicates values not recorded for the given run.