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

PREPARED BY: IPHC SECRETARIAT (B. HUTNICZAK; 8 MAY 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:

- Evaluating multiple components of the deep learning workflow for Pacific halibut age estimation, including model architectures, image preprocessing, weighting strategies, and training procedures.
- Evaluating the contribution of auxiliary biological and environmental covariates, through comparison of image-only and multi-input deep learning models integrating both visual and tabular data.
- Evaluating model temporal generalization by comparing age predictions from a model trained on images from one year to those from a different year.
- Evaluating model spatial generalization by comparing age predictions from a model trained on images from one area to those from a different area.
- Utilizing confidence intervals derived from deep ensemble techniques to assess the model's capability in identifying ambiguous or noisy samples.
- Demonstrating improvements in model performance associated with expansion of the image database from 2,799 (SRB026) to 11,107 otolith images.

UPDATES SUMMARY

Expanding image database

The IPHC continues to expand its otolith image database to support development and evaluation of deep learning approaches for Pacific halibut age estimation. Since the previous SRB meeting, the number of images available for model development has increased substantially, from 2,799 images to 11,107 otolith images used in the primary model runs. This increase in dataset size has improved representation across age classes, while also supporting more robust model training, evaluation, and testing of additional preprocessing and modelling approaches.

Cleaning pipeline

To evaluate whether removal of background information improves age prediction performance, a semi-automated image cleaning pipeline based on otolith segmentation was developed. First, a convolutional neural network with a U-Net architecture was trained to segment otolith regions from raw images using a limited set of manually annotated masks (approximately 200 images). The model was trained using a combination of binary cross-entropy and Dice loss to optimize both pixel-wise accuracy and region overlap, and standard data augmentation (including flips, rotations, and brightness variation) was applied to improve generalization.

Predicted segmentation masks were subsequently reviewed and corrected using a custom graphical interface, allowing manual refinement and exclusion of low-quality images. The final masks were then used to generate cleaned images by removing background pixels (set to white) and centering the otolith within the frame. As part of preprocessing experiments, grayscale conversion and contrast enhancement (CLAHE) were optionally applied within the segmented region. Images were finally resized to match the input resolution required by the age prediction model. This pipeline ensured that the downstream model was trained on images containing only the biologically relevant otolith structure, enabling a controlled comparison between raw and background-removed inputs.

Use of covariates

To assess the contribution of auxiliary biological and environmental information to model performance, parallel model variants were evaluated using identical image inputs. One model was trained using image data only, while a second model incorporated additional covariates alongside image features within a multi-input architecture. The covariates included geographic coordinates of capture (latitude and longitude), date of capture (expressed as day of year), and fish sex, where available. These variables were standardized prior to model input and processed through a dedicated fully connected branch, which was subsequently concatenated with features extracted from the convolutional backbone before final prediction. This design enabled joint learning from visual and tabular data while maintaining comparability between model variants. Performance differences between the image-only and image-plus-covariates models were therefore attributable to the inclusion of auxiliary information, allowing a direct evaluation of whether non-image features improved age prediction accuracy under otherwise identical training conditions.

Weights

To account for variability in the reliability of age determinations, a weighting scheme was applied during model training. Each image was assigned a weight proportional to the number of independent age readings conducted for the corresponding otolith, reflecting the confidence in the final age estimate. Higher weights were therefore assigned to samples with multiple consistent readings, while lower weights were applied to those based on fewer observations. This approach ensured that the loss function placed greater emphasis on more reliable labels, aligning model training with the underlying uncertainty in the age determination process.

Treating systematic negative bias at older age as an imbalanced-regression problem

Systematic underestimation observed in older age classes was addressed by testing an imbalanced-regression framework. Specifically, label distribution smoothing (LDS) was implemented following Yang et al. (2021), in which the empirical age distribution is smoothed using a Gaussian kernel to estimate an effective label density. Training samples are then reweighted inversely to this density, increasing the contribution of underrepresented older ages during model training.

In addition to LDS, feature distribution smoothing (FDS), as described by Yang et al. (2021), may be considered in future work for addressing imbalance. Unlike LDS, which operates on the label distribution, FDS aims to improve representation learning by smoothing feature statistics across adjacent target values during training. This is achieved by aligning feature representations of samples with similar target values using smoothed estimates of feature means and variances, thereby reducing noise and instability in sparsely populated regions of the target space.

Technical improvements

The latest model runs include a number of technical improvements. Image preprocessing was standardized through the use of model-specific normalization functions rather than simple scaling, ensuring compatibility with pretrained architectures. Additionally, data handling was refined by introducing seed-controlled shuffling and train–validation splitting, improving reproducibility across runs. The validation pipeline was also corrected by removing augmentation from the validation generator and disabling shuffling, allowing for more reliable performance assessment.

Testing spatial generalization

To evaluate the spatial generalization capability of the model, a targeted hold-out experiment was conducted using IPHC Regulatory Area as the spatial grouping variable. For each IPHC Regulatory Area, a fixed subset of 100 images was randomly selected and reserved as an independent test set. Two training scenarios were then implemented: (1) a spatial exclusion model, in which 2000 images excluding those originating from the focal IPHC Regulatory Area were randomly drawn for the training and validation sets, and (2) a control model, in which an equivalent number of images from all areas were randomly drawn for the training and validation sets to match sample size. Model performance was assessed on the held-out test subset using standard metrics including RMSE, MAE, R^2, and classification-based accuracy measures (e.g., exact match and within-one-year agreement). This design allowed for direct comparison between models trained with and without spatial representation of the test region, providing a structured assessment of the model's ability to generalize across geographic areas.

Latest results summary

Results:

1. In general, the use of covariates improved the predictive performance of the model, although the gain was relatively minor. Runs that included sex, coordinates, and date performed similarly to runs without sex, and in some cases resulted in lower MAE. Therefore, sex was not included in further testing, and variables that are readily available at the time of collection were prioritized.
2. Across RMSE, MAE, and R^2 metrics, standard images served as better inputs for training. Background cleaning, in some cases, resulted in a marginal increase in the percentage of predictions within a one-year tolerance, but grayscale conversion generally led to worse performance. Cleaned images, however, resulted in substantially faster training (24–30% reduction in runtime for runs with covariates).
3. Fine-tuning continues to provide substantial improvement to model generalization. For the selected primary run, predicting age for 2024 images using a model trained on 2019 otolith images resulted in a 24% higher RMSE. In contrast, when the model was fine-tuned on 20% of new images selected based on deep ensemble cross-validation, RMSE was only marginally higher (3%).

Results for individual model runs are available in Appendix B: Selection of model runs. Detailed evaluation and performance metrics for the selected best-performing model specification are presented in Preliminary results section and Appendix C: Deep ensemble individual results.

Spatial generalization results

The results of the spatial generalization experiment indicate that model performance is generally robust across Regulatory Areas, but the effects of excluding spatial information are not entirely consistent. In some areas (e.g., 2C, 4D, and, to a lesser extent, 3B), the control model that

included data from the focal Regulatory Area in training produced slightly higher predicted ages (Figure 1) and, in many cases, marginally lower MAE (Figure 2), suggesting a modest benefit from incorporating local data. However, this pattern was not universal: in some regions (notably 2A and 3B), the spatial exclusion model performed slightly better, and differences between models were often small relative to the variability observed across training seeds. The seed-level distributions further highlight that between-run variability can overlap substantially with the effect of spatial inclusion, indicating that model uncertainty is of similar magnitude to the spatial effect itself.

This interpretation is limited by the experimental design. Only three random seeds were used per model–area combination, reflecting the large number of total runs required across nine Regulatory Areas[1] and two training scenarios. As a result, estimates of variability and central tendency are based on a limited number of realizations, and some of the observed inconsistencies may reflect stochastic variation rather than systematic spatial effects. Nonetheless, both models tended to preserve overall age structure across regions, with mean predictions remaining close to observed values, supporting the conclusion that the model captures broadly transferable otolith features. Taken together, these findings suggest that while there is some benefit to including Regulatory Area–specific data, the model demonstrates a reasonable capacity for spatial generalization, and performance degradation under complete spatial exclusion is limited and not consistently biased across all areas.

It is noted, however, that all IPHC Regulatory Areas are currently represented in the available dataset, and therefore the experimental design reflects a hypothetical scenario in which data from a given Regulatory Area are entirely absent. As such, the results should be interpreted as an assessment of potential model performance under conditions of incomplete spatial coverage, rather than a direct reflection of current operational limitations. In contrast, temporal generalization represents a more critical consideration, as models would ideally be available for routine application to new-year data for which no contemporaneous training data are available.

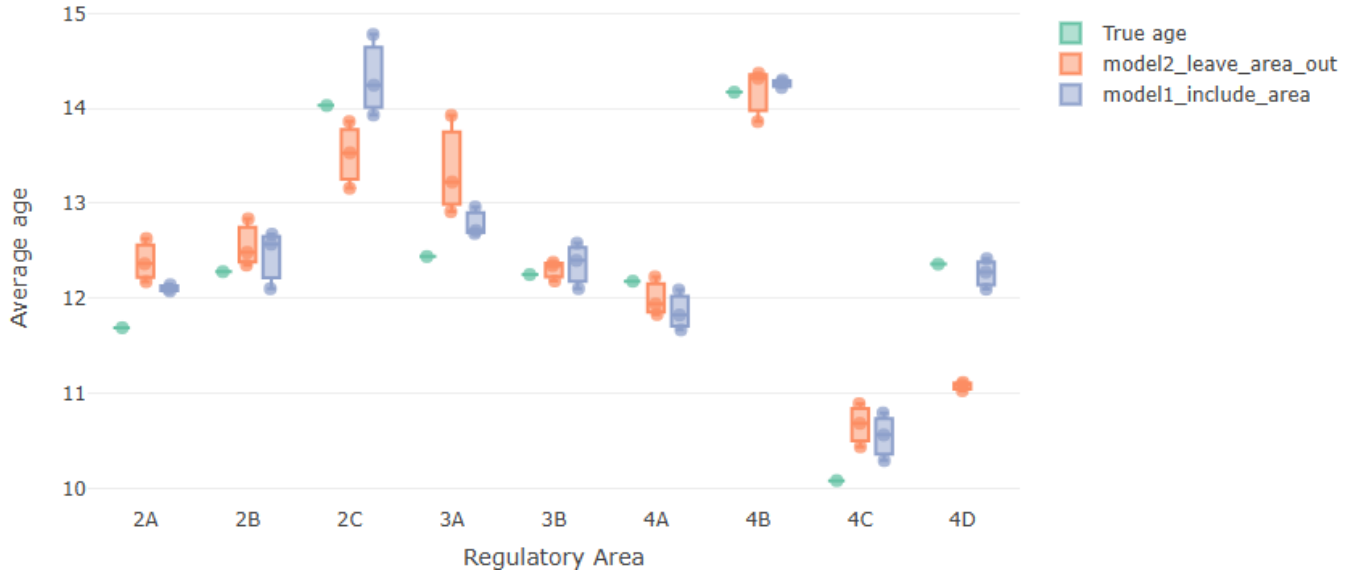

Figure 1: Spatial generalization results - manually derived age vs. model predictions (seed-level).

[1] Training data were not available for Regulatory Area 4E.

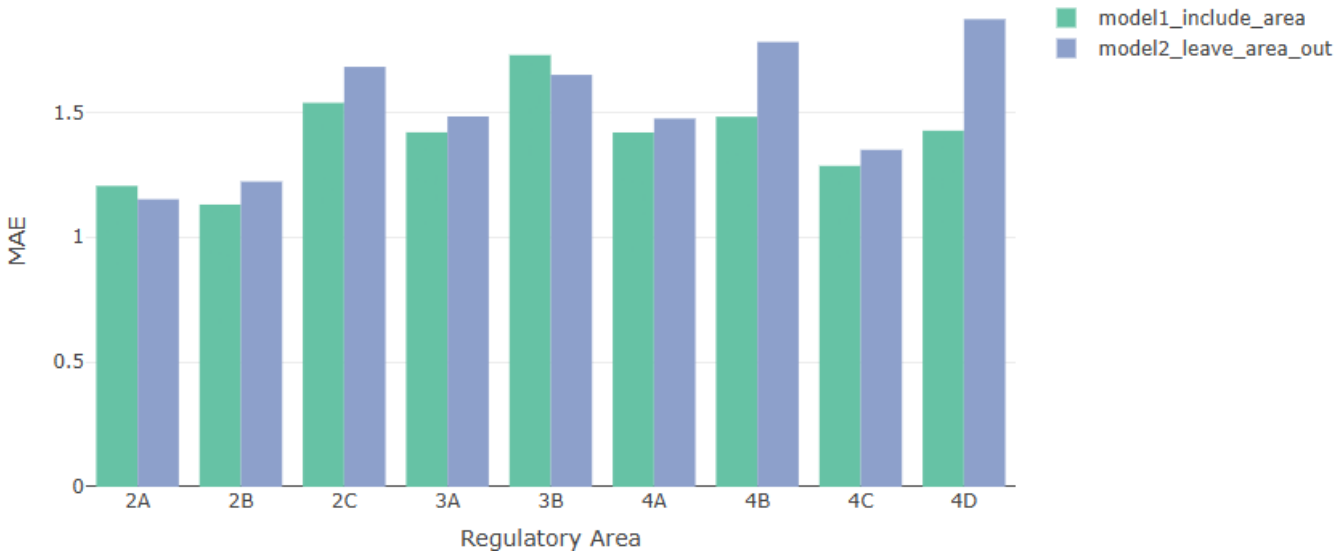

Figure 2: Spatial generalization results - mean average error (MAE) by Regulatory Area and model.

Evaluation of Label Distribution Smoothing

A model version incorporating label distribution smoothing (LDS) was evaluated in an attempt to reduce the tendency of the model to underpredict older Pacific halibut ages. The approach increased the weighting assigned to sparsely represented age classes, effectively encouraging the network to prioritize accurate prediction of relatively few older individuals during training. However, this implementation substantially reduced overall predictive performance, with test-set RMSE increasing to 5.1725 compared to 1.7921 for the standard setup using covariates. The results suggest that the LDS weighting strategy over-corrected for age imbalance, causing the model to fit a limited number of difficult and potentially noisy older-age samples at the expense of broader generalization across the dataset. Given the comparatively small sample sizes and greater uncertainty typically associated with older otolith ages, the weighting scheme likely amplified noise rather than biological signal. Consequently, the resulting performance fell outside the range considered operationally useful for ageing. Although the present implementation was unsuccessful, additional investigation of imbalance-aware learning approaches, including less aggressive weighting schemes or alternative calibration strategies, may still be warranted in future work.

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,[2] and can be subject to bias[3] as well as imprecision due to variations in age estimations between readers and within readers over time. Recent advances in imaging technologies and machine learning suggest that AI can assist in this process by automating the analysis of otolith images[4] 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

Model framework

The proposed model framework (Figure *3*) includes a continuous process of training the model using available labelled data (aged otoliths), querying the model to select the next sample, labeling or relabeling the selected sample, and enriching the model with newly labelled samples.

This model relies on automatized ageing that is supplementing the expert-derived age estimates continuously improving the model in the *Label* phase and the *Enrich* phase.

[2] 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.

[3] 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 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.

[4] 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).

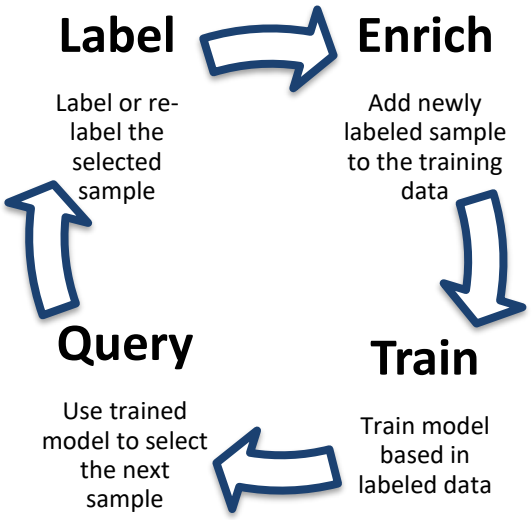

Figure 3. Model framework.

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*).[5]

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.

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

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

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 TensorFlow model development.

The approach utilizes a transfer-learning technique to develop a CNN for otolith age estimation. Transfer learning is the process of repurposing a machine learning model that has been pre-trained for another, related, task. Specifically, it starts with the InceptionV3 model from Google, pre-trained on the ImageNet database. ImageNet database contains over 14 million annotated images classified into 1,000 categories. By loading CNN layers with publicly available pre-trained weights rather than random initialization, transfer learning significantly enhances model performance.

To adapt this model specifically for Pacific halibut ageing, modifications included scaling the input layer to match otolith images' resolution[6] and changing the output from multi-dimensional class probabilities to a single numeric output for regression.[7] Thus, the architecture employed follows the pattern: Input → InceptionV3 (feature extractor) → Regressor → Output, optimized

[6] Resolution is the total number of pixels along an image's width and height, expressed as pixels per inch (PPI). The Inception v3 model processes images that are 299 x 299 pixels in size. The original images (2548 × 2548 pixels) were first resized to 400 × 400 pixels prior to input into the model. This intermediate resizing step preserves more visual detail than a direct downscaling to 299 × 299 and allows for subsequent data augmentation operations (such as cropping, flipping, or rotation) to be applied more effectively before the final resize to the model's required input size.

[7] 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.

using stochastic gradient descent (SGD) to minimize mean squared error (MSE) between model predictions and expert annotations.[8]

A similar approach, although adopting classification approach, was applied for ageing Greek Red Mullet (*Mullus barbatus*) (Politikos et al., 2022) and the associated code is available on GitHub (github.com/dimpolitik/DeepOtolith). The available open-source code was adapted to test the approach for Pacific halibut.

In addition to the InceptionV3 architecture, alternative architectures continue to be explored to identify potential improvements in predictive performance and computational efficiency. These include EfficientNet variants (EfficientNetB4, EfficientNetB5, EfficientNetV2 M/L) and ConvNeXt. EfficientNet architectures are known for their balanced approach to scaling depth, width, and resolution, optimizing computational efficiency and accuracy. EfficientNetV2 further refines this by introducing progressive training and improved scaling techniques. ConvNeXt architectures, inspired by transformer models, incorporate modifications to convolutional structures, achieving competitive accuracy with a simplified design and potentially improved model interpretability.

However, given previously reported underperformance of alternative architectures (IPHC-2025-SRB026-10), the current update focuses primarily on results derived from the InceptionV3 implementation. Despite their advanced theoretical advantages - such as better scalability, computational efficiency, and deeper learning capabilities - EfficientNet and ConvNeXt models underperformed relative to the simpler InceptionV3 architecture. Several configurations of EfficientNet and ConvNeXt exhibited limited learning, with predictions regressing toward the mean age of the test dataset. This outcome suggests that these more complex models struggled to extract meaningful age-related features from the otolith images, likely due to a combination of insufficient training data and overfitting driven by model complexity.

In contrast, the InceptionV3 architecture consistently produced more accurate and stable predictions, indicating that its comparatively simpler structure may currently be better suited to the available dataset size and image variability. Furthermore, the improvements achieved through additional refinements presented in this update suggest that there remains considerable opportunity for further optimization within the InceptionV3 framework itself.

TensorFlow/Keras has been the primary framework used in the current implementation. However, future work may explore alternative frameworks such as PyTorch (originally developed by Meta), which offers flexible dynamic computation graphs and growing adoption in the deep learning research community.

Performance metrics and achieved accuracy

Performance of the CNN to correctly assign ages (rounded output of the regression layer) to otolith images in the test set is assessed via the root mean squared error (RMSE), mean average error (MAE) and the percentage of correctly predicted ages, as well as predictions within ±1 year tolerance. Moen et al., (2018) also suggest calculating coefficient of variation (CV).[9]

Moen et al., (2018), for Greenland halibut, achieved MSE for the left and right otoliths and pair of 3.27, 2.71 and 2.99, respectively. Age was correctly estimated for 48 out of the 164 tested otolith-pairs (29%). In addition, 63 cases (38%) were estimated to be one year off the read age.

[8] In practice, the neural network minimizes the MSE of normalized age values, i.e., age values divided by the maximum age provided as input.

[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).

There was also a clear tendency for the system to predict a lower age for older individuals, when compared to human readers. The variance of the predictions also increased with the age of the otolith.

The model developed by Moore et al. (2019), for prediction of age of snapper using CT scans,[10] gave the same age as the human reader for 47% of otoliths in a test dataset, 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 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 of otoliths, supplemented by geospatial and biological data routinely collected on the survey, estimated age of walleye pollock. For the optimal multimodal CNN model, an RMSE of 0.83 for the training set and an RMSE of 0.91 for the test set indicated that at least 67% of estimated ages were predicted within ±1 year of age compared to traditional microscope-based ages.

However, it should be noted that neither the traditional ageing methods for Pacific halibut are perfectly accurate. Within- and between-reader agreement in age assignment is generally 60%-70% complete agreement, 80% to 90% within one year, and 100% within 3 years. The IPHC Secretariat's publications report on % agreement (see *Technical Report No. 46* and *No. 47*).

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]

[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.

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

[12] Moen et al. (2018) used images 400 by 400 pixels, which required the input layer to be scaled to match the Inception V3 requirements (299 by 299 pixels). Ordoñez et al. (2020), using the same set of images, built a CNN with images resized to 224 by 224 pixels, the default input of the VGG-19 model. 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.

It is important to note that 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 (Moore et al., 2019).

Figure *4* shows an example of a range of images used in the CNN training dataset.

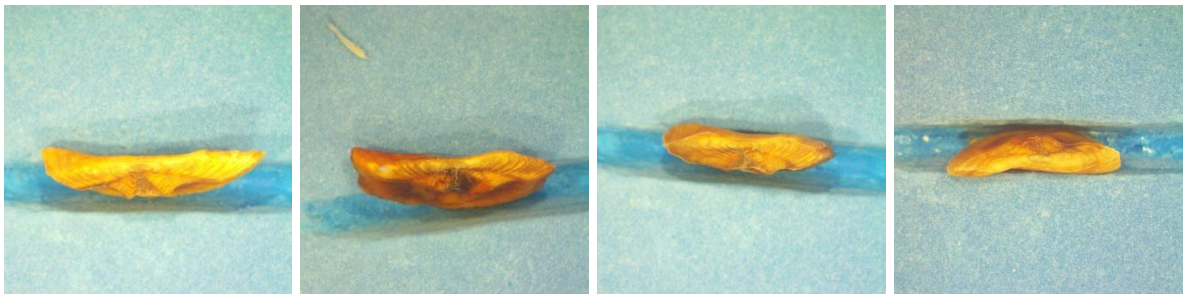

Figure 4. Examples of Pacific halibut otolith images taken for inclusion in the training set.

In addition, the IPHC is in the process of creating complementary database comprising labelled images of otoliths captured prior to processing to conduct a cost-benefit analysis of using processed versus unprocessed otoliths for AI-based age determination. Example images are provided in Figure *5*. In their research, Politikos et al. (2022) utilized digital images of otoliths that were not subject to any additional processing in the laboratory, immersed in water and placed under a stereomicroscope on a white background with transmitted light. However, it is important to note that even if results indicate that breaking and baking is not necessary for age determination using AI, a subsample chosen for the Label and Enrich phases would have to be fully processed for age determination with traditional methods by an expert reader.

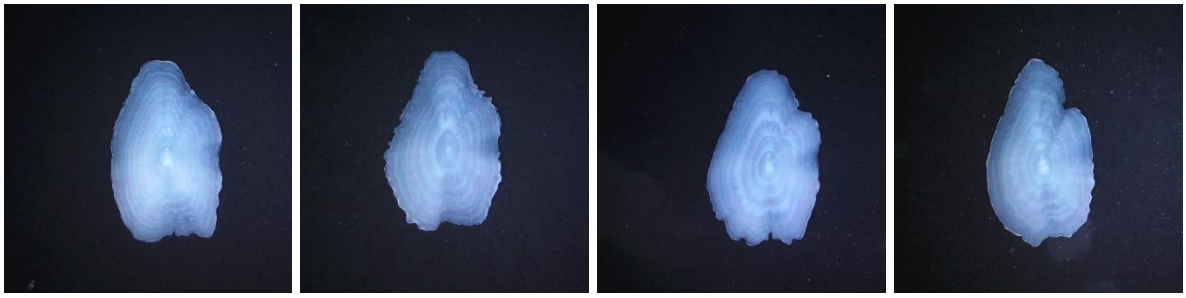

Figure 5. Examples of Pacific halibut otolith images taken for inclusion in the training set.

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[13] are omitted when collecting samples.

Ongoing research [Dimitris Politikos, personal communication] is investigating the initial stage of the aging process, specifically assessing whether an otolith is of sufficient quality for age determination. This research is relevant for cases involving crystallized or broken otoliths and aims to potentially eliminate the need for subjective decisions by samplers regarding the usability of otoliths for age determination. This approach implements a two-stage classification system. In the first stage, the model assesses the otolith's suitability for ageing; in the second, it

[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.

determines the age. Th algorithm-driven presorting could also incorporate expert knowledge for handling problematic otoliths.

In developing the model, the training dataset can be strategically supplemented with images of samples that represent a group of otoliths with which the original model struggles the most (Query phase).[14]

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. Year collected – to account for variation between cohorts and prevalent environmental conditions;[15]
5. Date collected – to account for the 'edge effect' reflecting seasonal changes;
6. Geospatial characteristics of the collection site (latitude, longitude and IPHC Regulatory Area) – to capture regional variation;
7. Resolved sex – to determine whether otolith characteristics (possibly not directly visible to human eye) could be used for sex determination.[16]

Uncertainty estimates

To further refine accuracy in a production setting, a mixed-method approach can be applied. This approach involves selecting a subset of otolith images (e.g., 10%, 20% or 50%) for ageing by human experts, with selection guided by model-derived uncertainty estimates. Under this framework, images associated with high predictive uncertainty would be prioritized for expert review, while high-confidence predictions could be processed automatically. Newly validated samples could subsequently be incorporated into the training dataset for annual fine-tuning, enabling targeted and resource-efficient model improvement over time.

In practice, this strategy would allow human experts to focus on "difficult" otoliths, while automating the processing of comparatively "easy" samples with high model confidence. Such a hybrid workflow has the potential to improve throughput without compromising the accuracy and consistency necessary for applications such as stock assessment, where minimizing systematic bias is critical.[17]

Several approaches were considered for quantifying model uncertainty, including Monte Carlo dropout (Gal & Ghahramani, 2016) and deep ensemble method (Lakshminarayanan et al., 2017). Following preliminary 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

[14] About 1% of otoliths are partly crystallized and are assigned ages. The same is true for broken otoliths that are aged (1%)

[15] 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.

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

[17] If there is a strong junction in the relative precision between old and younger fish due to the change in methods this may require a nonparametric approach to ageing imprecision. If an AI method is biased as a function of age (standard for surface reading methods) and the break and bake method is unbiased, integrating the methods may prove challenging.

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

PRELIMINARY RESULTS

Selected model evaluation

The selected model configuration 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. Additionally, a separate set of 2,931 images of otoliths collected during the 2024 FISS was used to verify model performance on additional unseen data, testing the temporal generalization of the model configurations. All images were resized to 400x400 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 5 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 C*.

Across ensemble runs, the model trained for an average of 183 epochs (123 effective epochs with early stopping set at 60). It achieved a normalized MSE average of 0.00114 on the validation set. When averaged across seeds results were rounded to the nearest integer age, the RMSE for the test set was 1.68 and MAE was 1.06. On average, the ensemble predicted the exact age correctly for 35.5% of test images, and an additional 41.0% were within ±1 year of the manually assigned age, resulting in a total agreement within 1 year for 76.5% of cases.

Figure *6* shows a comparison between manually derived ages and AI-predicted ages across the ensemble. Figure 7 compares the age composition estimated manually with that derived from the ensemble model predictions.

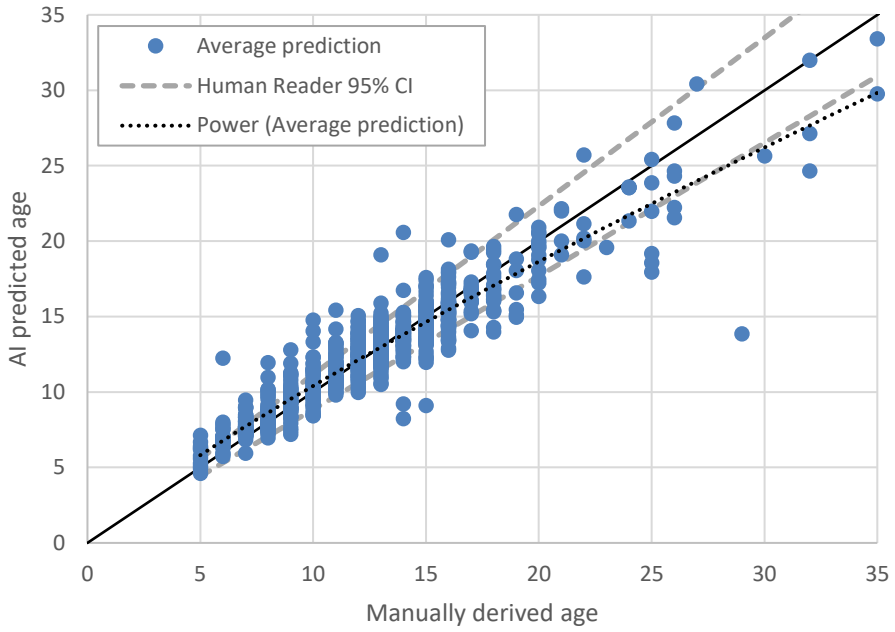

Figure 6. Comparison between manually derived age with AI predicted age.

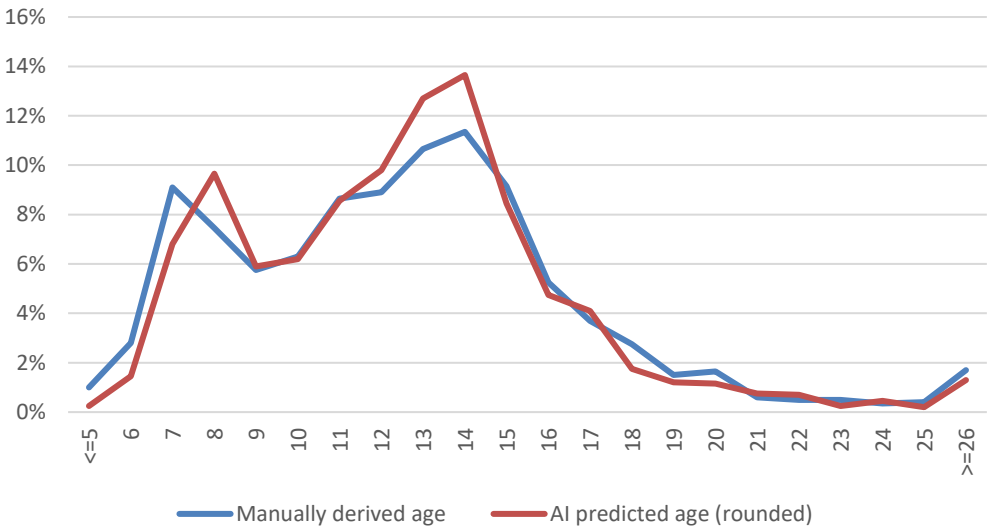

Figure 7. Comparison between manually derived age with AI predicted age – age composition.

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 *8*). 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.

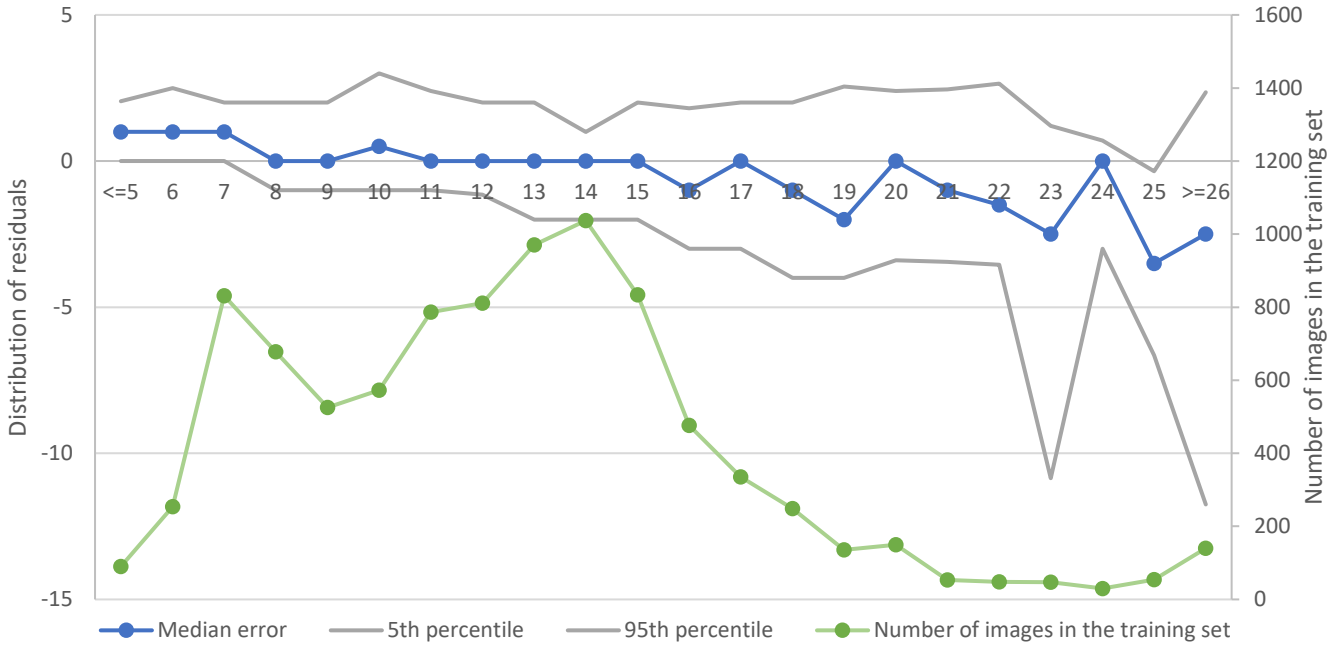

Figure 8. Distribution on residuals and number of images by age in the test set.

Testing temporal generalization

The performance of the model trained on the 2019 FISS sample declined when applied to otolith images collected during the 2024 survey, reflecting the challenges of temporal generalization. On average, the root mean squared error (RMSE) increased to 2.27, representing an approximate 35% increase compared to the 2019 test set. Furthermore, the proportion of predictions within ±1 year of the manually assigned age dropped by 16.9 percentage points, indicating a decline in predictive accuracy.

However, the use of a deep ensemble approach enabled a more nuanced evaluation of model reliability. Specifically, the ensemble framework provided per-sample uncertainty estimates (measured as the standard deviation across model predictions), which helped distinguish between confidently and less confidently predicted samples. This enabled stratification of predictions by uncertainty level.

Figure 9 shows the cumulative proportion of 2024 test samples for which the ensemble prediction falls within ±1 year of the manually assigned age, as a function of increasing prediction uncertainty (measured by the standard deviation across the ensemble). The curve confirms that predictions with lower uncertainty levels tend to be more accurate. For the least uncertain subset of the test data (e.g., the first ~20%), accuracy within ±1 year exceeds 80%, while this metric gradually declines as predictions with higher uncertainty are included. By the time the entire sample is considered, accuracy drops to approximately 59%.

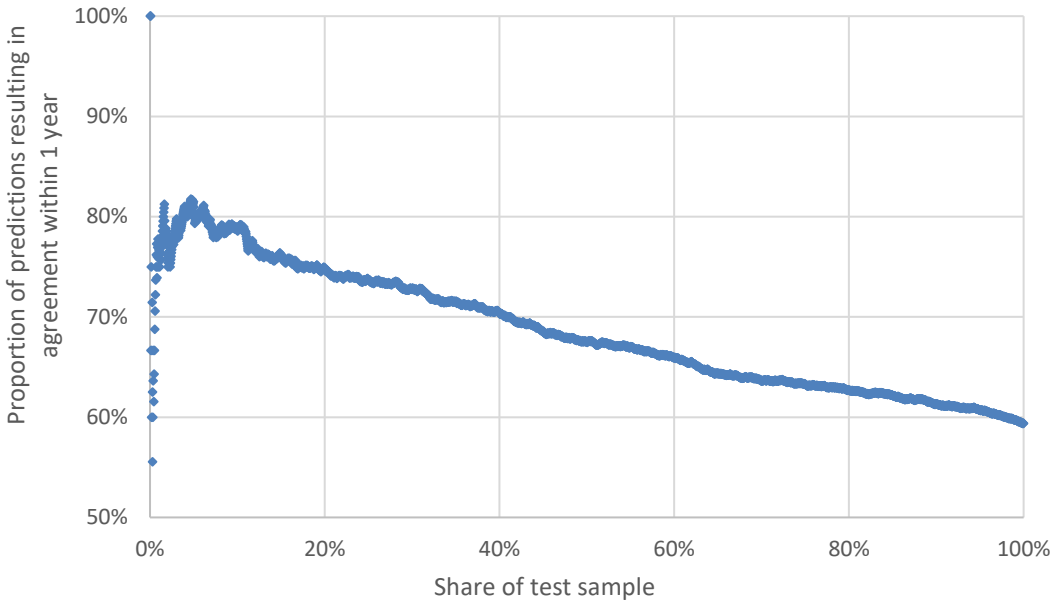

Figure 9: Proportion of ensemble predictions within ±1 year of manual age as a function of cumulative share of the test sample, ordered by prediction uncertainty (standard deviation).

Fine-tuning the model

To assess the impact of fine-tuning on model generalization across years, the ensemble originally trained on 2019 FISS images was fine-tuned using a randomly selected 20% subset of otoliths collected in 2024. The model was then evaluated on the remaining unseen 80% of 2024 images. Fine-tuning yielded measurable improvements: the average RMSE across ensemble runs decreased from 2.27 to 2.23, and the proportion of predictions within ±1 year of the manually assigned age increased from 59.6% to 66.1%.

In a separate analysis, the fine-tuning subset was selected based on uncertainty rather than random sampling. Specifically, 20% of 2024 images with the highest standard deviation across ensemble predictions - interpreted as the most ambiguous or noisy samples - were used for fine-tuning. This targeted approach led to further gains in predictive accuracy. When evaluated on the remaining 80%, the model achieved an RMSE of 1.97 and the proportion of predictions within ±1 year of the manually assigned age of 70.0%

CONCLUSIONS

The ongoing advancement of AI technologies in the field of marine science offers considerable potential to enhance the efficiency of age determination of Pacific halibut using otolith images. Preliminary results presented here suggest that convolutional neural networks (CNNs), particularly when implemented using a deep ensemble approach, could provide predictive accuracy that supports their use as a supplement - or in some cases, a potential alternative - to the current manual ageing protocol.

The results demonstrate that continued expansion of the image database improves model development and evaluation capacity. Increasing the training dataset from 2,799 to 11,107 otolith images improved representation across age classes and biological variability, supporting more robust model training and enabling evaluation of additional preprocessing, weighting, and modelling approaches. Incremental improvements were also observed through continued refinement of the modelling pipeline itself, including implementation of model-specific preprocessing, improved data handling procedures, and evaluation of auxiliary biological and

environmental covariates. Although the inclusion of covariates such as capture location and collection date resulted in comparatively modest gains in predictive performance, the results suggest that non-image information can contribute additional biologically relevant context to the model under otherwise identical training conditions.

The findings also highlight the practical value of the deep ensemble framework. In addition to improving predictive performance, ensemble-based models provide per-sample uncertainty estimates that can be used to identify potentially unreliable predictions. This enables a mixed-method protocol in which low-confidence predictions, identified through high variance across ensemble members, can be flagged for expert review, while high-confidence outputs may be accepted directly. Such an approach could substantially streamline the ageing workflow while maintaining the accuracy and consistency required for stock assessment applications.

Results additionally showed that model performance deteriorates when predictions are applied to data collected in years different from those represented in the training dataset, indicating limited temporal generalization. However, modest fine-tuning using a subset of current-year images substantially improved predictive performance, reducing RMSE and increasing agreement within ±1 year of expert-derived ages. Fine-tuning focused specifically on uncertain samples identified through ensemble variance produced further improvements, suggesting that uncertainty-guided updating may represent an effective strategy for adapting models to new data while minimizing the amount of manual ageing required.

Spatial generalization experiments indicated that the model retains a reasonable ability to generalize across IPHC Regulatory Areas, with only limited and inconsistent performance degradation when data from a focal area were excluded from training. Although inclusion of area-specific data provided modest improvements in some regions, the overall results suggest that the model captures broadly transferable otolith features and is not strongly dependent on spatially localized training data.

Despite the encouraging progress, important limitations remain. Although the latest model runs showed no statistically significant bias for the oldest age categories (~21+ years), the model still tends to underestimate ages of older Pacific halibut on average. These age classes remain substantially underrepresented within the available training data, reflecting the underlying population structure. About 4.1% of otoliths included in training of the main model were from fish aged 21 years or older. Attempts to directly correct this imbalance using label distribution smoothing resulted in substantially poorer overall predictive performance, suggesting that aggressive imbalance-correction approaches may amplify label noise and reduce broader model generalization. Expanding the dataset to improve representation of older individuals will therefore remain an important priority for achieving more balanced training and improving reliability across the full biological age range.

Finally, it is important to emphasize that AI-based ageing models will continue to rely on human experts, both for validation and for generation of the high-quality training data that reflect temporal changes and environmental variability. As environmental conditions and stock structure continue to change, integrating expert oversight and continual model updating will remain a critical part of accurate AI implementation for ageing process.

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

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: SELECTION OF MODEL RUNS

RunID	1	2	3	13	14	15	16	17	18	19	20	21	22	24
SETUP				**									**	
Architecture	InV3	InV3	InV3	InV3	InV3	InV3	InV3	InV3	InV3	InV3	InV3	InV3	InV3	InV3
Image mode	rgb	rgb clean	grayscale clean	rgb	rgb clean	grayscale clean	rgb	rgb clean	grayscale clean	rgb	rgb clean	grayscale clean	rgb	rgb
Covariates	-	-	-	s+c+d	s+c+d	s+c+d	-	-	-	c+d	c+d	c+d	c+d	c+d
Notes													technical improvements	LDS
RESULTS														
Validation MSE	0.0012	0.0013	0.0014	0.0011	0.0012	0.0012	0.0014	0.0012	0.0014	0.0011	0.0013	0.0013	0.0012	0.0121
Epochs trained	184	164	145	184	155	159	154	148	141	192	146	137	192	157
Test RMSE	0.00129	0.00136	0.00136	0.00125	0.00135	0.00152	0.00141	0.00132	0.00135	0.00126	0.00135	0.00148	0.00125	0.01138
Test MAE	0.0244	0.0247	0.0251	0.0245	0.0252	0.0275	0.0255	0.0243	0.0249	0.0240	0.0251	0.0268	0.0237	0.0778
Test R^2	0.84	0.83	0.83	0.85	0.83	0.81	0.83	0.84	0.83	0.84	0.83	0.82	0.85	-0.41
Correctly predicted	30.6%	29.9%	29.2%	29.6%	29.5%	26.5%	28.8%	31.5%	29.6%	31.0%	29.8%	27.0%	31.2%	9.7%
Correctly predicted with ±1 year tolerance	71.9%	72.3%	70.7%	72.0%	71.6%	67.1%	70.1%	72.2%	71.2%	72.8%	71.3%	67.6%	73.4%	28.0%
RUN parameters														
Run time in hours	15.0	12.1	10.9	15.4	11.8	11.9	11.3	10.8	10.6	15.8	11.0	10.4	15.6	12.8
RESULTS for 2024														
RMSE	*	*	*	*	*	*	*	*	*	2.24414	2.24892	2.36083	2.27066	*
MAE	*	*	*	*	*	*	*	*	*	1.5684	1.5660	1.6841	1.5844	*
R^2	*	*	*	*	*	*	*	*	*	0.795	0.793	0.772	0.791	*
Correctly predicted	*	*	*	*	*	*	*	*	*	21.9%	23.3%	20.2%	21.8%	*
Correctly predicted with ±1 year tolerance	*	*	*	*	*	*	*	*	*	60.3%	59.8%	55.3%	59.6%	*
RESULTS for 2024 (FT)														
RMSE	*	*	*	*	*	*	*	*	*	2.21941	2.22268	2.26534	2.22613	*
MAE	*	*	*	*	*	*	*	*	*	1.4614	1.4426	1.4832	1.4358	*
R^2	*	*	*	*	*	*	*	*	*	0.808	0.806	0.798	0.807	*
Correctly predicted	*	*	*	*	*	*	*	*	*	25.2%	27.6%	25.9%	26.7%	*
Correctly predicted with ±1 year tolerance	*	*	*	*	*	*	*	*	*	64.9%	65.8%	64.4%	66.1%	*
RESULTS for 2024 (FT select)														
RMSE	*	*	*	*	*	*	*	*	*	1.86369	1.98599	1.97036	1.96559	*
MAE	*	*	*	*	*	*	*	*	*	1.3488	1.3761	1.4064	1.3510	*
R^2	*	*	*	*	*	*	*	*	*	0.820	0.794	0.791	0.812	*
Correctly predicted	*	*	*	*	*	*	*	*	*	22.9%	25.0%	22.7%	24.2%	*
Correctly predicted with ±1 year tolerance	*	*	*	*	*	*	*	*	*	66.1%	65.9%	63.7%	67.0%	*

Note: Results represent averages across three individual runs using randomly selected seed values; individual run performance varied. All models utilized the InceptionV3 (InV3) architecture, with image size = 400 × 400 pixels, dropout = 0.25, and L2 regularization = 1.0×10^{-4}. Training parameters included a maximum of 600 epochs, batch size = 8, EarlyStopping patience = 60, ReduceLROnPlateau patience = 30 epochs with reduction factor = 0.8, and an initial learning rate = 0.0002. Full augmentation settings included rotation range = 360°, width shift range = 0.1, height shift range = 0.1, brightness range = [0.95, 1.05], and zoom range = [0.98, 1.02]. Covariate legend: c = coordinates, d = date collected, s = sex.

Machine setup: VM with AMD EPYC 7V12 64-Core Processor and Nvidia Tesla T4 GPU.

* Indicates values not recorded for the given run.

**Indicates model selected for further investigation.

APPENDIX C: DEEP ENSEMBLE INDIVIDUAL RESULTS

Model run	1	2	3	4	5	**AVERAGE**
Seed	11	27	42	44	51	
Epochs trained	219	156	200	194	147	183
Validation MSE	0.00106	0.00111	0.00131	0.00113	0.00111	0.00114
Rum time [h]	18.0	12.6	16.1	15.7	11.9	14.9
RESULTS – TEST SET						
Test RMSE	1.814	1.782	1.780	1.802	1.830	1.802
Test MAE	1.161	1.182	1.145	1.172	1.193	1.170
Test R^2	0.841	0.848	0.847	0.844	0.839	0.844
Correctly predicted	32.0%	29.9%	31.7%	31.9%	30.5%	31.2%
Correctly predicted with ±1 year tolerance	73.6%	72.3%	74.2%	72.7%	72.8%	73.1%
RESULTS – 2024 IMAGES						
RMSE	2.466	2.418	2.361	2.426	2.513	2.437
MAE	1.734	1.678	1.661	1.720	1.754	1.709
R^2	0.751	0.759	0.772	0.760	0.757	0.760
Correctly predicted	20.8%	22.0%	20.7%	20.6%	20.7%	21.0%
Correctly predicted with ±1 year tolerance	56.0%	58.4%	57.3%	55.8%	57.1%	56.9%
RESULTS – 2024 IMAGES – 20% FT						
RMSE	2.392	2.330	2.295	2.333	*	2.337
MAE	1.581	1.538	1.489	1.593	*	1.550
R^2	0.775	0.786	0.793	0.786	*	0.785
Correctly predicted	24.6%	25.7%	26.1%	22.6%	*	24.7%
Correctly predicted with ±1 year tolerance	61.5%	62.6%	65.1%	60.1%	*	62.3%
RESULTS – 2024 IMAGES – 20% FT select						
RMSE	2.027	2.046	1.966	*	*	2.013
MAE	1.414	1.444	1.351	*	*	1.403
R^2	0.798	0.796	0.812	*	*	0.802
Correctly predicted	23.0%	21.9%	24.2%	*	*	23.0%
Correctly predicted with ±1 year tolerance	63.2%	62.6%	67.0%	*	*	64.2%

* Indicates values not recorded for the given run.