

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

PREPARED BY: IPHC SECRETARIAT (B. HUTNICZAK, J. FORSBERG, K. SAWYER VAN VLECK, & K. MAGRANE; 5 MAY 2025)

PURPOSE

This document summarizes the information available on the use of artificial intelligence (AI) for determining the age of fish from images of collected otoliths and provides an update on the exploratory work of implementing an AI-based age determination model for Pacific halibut.

The progress summarized in this document includes:

- Testing various deep learning architectures to identify the optimal approach given the available otolith images.
- Evaluating model generalization by comparing age predictions from a model trained on images from one year to those from a different year.
- Assessing differences in model performance between images of processed (sectioned and baked) and unprocessed (surface) otoliths.
- Utilizing confidence intervals derived from deep ensemble techniques to assess the model's capability in identifying ambiguous or noisy samples.
- Evaluating the model's performance in predicting the geographic region of sample collection.

The purpose of this document is twofold. First, it provides essential background information to support ongoing efforts in establishing a comprehensive database of otolith images with expertprovided labels for future ageing use. Second, it provides an update on the viability of an AIbased modeling approach for supplementing current Pacific halibut ageing protocol, while also outlining the remaining steps and requirements necessary for operational implementation.

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,¹ and can be subject to bias² 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 images³ 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 1) 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.

¹ 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.

² 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.

³ 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 1. 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*).⁴

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

⁴ 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 architectural 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 pretrained for another, related, task. Specifically, it starts with the <u>InceptionV3 model from Google</u>, pre-trained on the <u>ImageNet database</u>. 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' resolution⁵ and changing the output from multi-dimensional class probabilities to a single numeric output for regression.⁶ Thus, the architecture employed follows the pattern: Input \rightarrow InceptionV3 (feature extractor) \rightarrow Regressor \rightarrow Output, optimized

⁵ 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.

⁶ 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.⁷

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 (<u>github.com/dimpolitik/DeepOtolith</u>). The available open-source code was adapted to test the approach for Pacific halibut.

In addition to the InceptionV3 architecture, alternative architectures were explored to identify potentially superior performance or efficiency advantages. These included EfficientNet variants (EfficientNetB4, EfficientNetB5, EfficientNetV2 S/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.

While TensorFlow/Keras has been the primary framework used in the current implementation, 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) 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).⁸

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. 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,⁹ 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

⁷ In practice, the neural network minimizes the MSE of normalized age values, i.e., age values divided by the maximum age provided as input.

⁸ 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).

⁹ 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.

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 <u>Technical Report No. 46</u> and <u>No. 47</u>).

Use of auxiliary data

The accuracy and precision of age predictions from otolith images using neural networks could potentially be enhanced by incorporating auxiliary data into the modeling process (Moen et al., 2018). For example, the geographic location where fish are captured could offer valuable supplementary information to the model. Past IPHC work suggests a good deal of spatial variation in Pacific halibut growth ring patterns. This points to the importance of good spatial coverage in the training sample.

The project plans to explore the integration of spatial covariates, such as latitude, longitude, or defined regulatory areas, to refine age predictions. Inclusion of these spatial factors could help the neural networks better interpret and account for region-specific growth patterns that influence otolith formation. Other available auxiliary data include collection year, which could be applied to account for variation between cohorts and prevalent environmental conditions throughout the aged fish life histories, and the collection dates, which provide insights into seasonal variation to the interpretation of the otolith edge.

Database

The IPHC annually ages a considerable number of otoliths (see <u>Appendix A</u> for details). Since 1925, over 1.5 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,¹⁰ 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.¹¹

¹⁰ The camera fits in one of the microscope eyepieces, eliminating the need to purchase a separate camera mount for the microscope.

¹¹ 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 2 shows an example of a range of images used in the CNN training dataset.





In addition, the IPHC is in the process of creating complimentary 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 3. 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.





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 otoliths¹² 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

¹² 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).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. Year collected to account for variation between cohorts and prevalent environmental conditions;
- 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.¹⁴

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% or 20 % - for reexamination by human experts, focusing specifically on cases where the AI model expresses low confidence in its predictions. These selections would be guided by model-derived uncertainty estimates. The newly relabeled samples can then be incorporated into the training set for annual fine-tuning, contributing to ongoing model improvement in a resource-efficient and targeted manner.

In practice, this strategy would allow human experts to focus on "difficult" otoliths—those with high uncertainty—while automating the processing of "easy" ones with high model confidence. This hybrid workflow enhances throughput without compromising the accuracy and consistency necessary for applications such as stock assessment, where minimizing systematic bias is critical.¹⁵

Two approaches were considered for quantifying model uncertainty:

• **Monte Carlo dropout** (Gal & Ghahramani, 2016): This technique involves performing multiple forward passes through the model with dropout layers activated during inference. The resulting variability in predictions across passes is used to estimate confidence intervals. Monte Carlo Dropout is computationally efficient and easy to implement, and it provides a useful proxy for identifying ambiguous or noisy samples. This form of persample uncertainty is also referred to as training dynamics or soft loss tracing.

¹³ About 1% of otoliths are partly crystallized and are assigned ages. The same is true for broken otoliths that are aged (1%)

¹⁴ IPHC is currently using genotyping for Pacific halibut sex determination.

¹⁵ 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.

• **Deep ensembles** (Lakshminarayanan et al., 2017): This approach involves training multiple independently initialized models and aggregating their predictions to form a consensus output. The variance across ensemble members serves as an estimate of prediction uncertainty. Deep ensembles are generally more robust than Monte Carlo Dropout, especially in identifying out-of-distribution samples and capturing both model and data uncertainty. Their main advantage lies in their improved predictive performance and better-calibrated confidence intervals, though at the cost of increased computational resources.

Together, these tools support the design of a semi-automated, quality-controlled ageing protocol that leverages the strengths of both AI and human expertise.

PRELIMINARY RESULTS

Comparison of model architectures

Several modern CNN architectures were systematically evaluated to determine the most suitable approach for ageing Pacific halibut using otolith images. The architectures tested included:

- InceptionV3: A widely used CNN known for its balanced computational efficiency and accuracy.
- EfficientNet (B4, B5, V2 S/M/L): Architectures optimized for scaling model depth, width, and resolution uniformly, enhancing computational efficiency and predictive accuracy.
- **ConvNeXt**: Inspired by transformer-based models, ConvNeXt utilizes modified convolutional operations aiming to simplify model complexity while maintaining competitive performance.

Each architecture was adapted via transfer learning, leveraging publicly available pre-trained weights from the ImageNet database, and subsequently fine-tuned specifically for the task of Pacific halibut age prediction. Adaptations involved resizing input images to match each architecture's requirements and adjusting the output layer to perform regression predicting age as a continuous numeric value.

The models were evaluated using standardized procedures to ensure valid and robust comparisons. The main evaluation criteria included:

- RMSE, percentage of exact age matches, and percentage within ±1 year tolerance between predicted ages and expert-provided ages for a test set of images collected within the same year as those used for training (without image overlap).
- RMSE, percentage of exact age matches, and percentage within ±1 year tolerance for a second test set comprising images collected five years after the training images, providing an assessment of temporal generalization.

The evaluation involved multiple experimental runs to ensure robustness. Selection of model run configurations and evaluation results are provided in <u>Appendix 2</u>.

The comparative evaluation revealed significant performance differences among tested CNN architectures. 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 derived more accurate and reliable predictions, suggesting that its simpler structure is more suitable given the current limitations in dataset size and variability. However, the selected final InceptionV3 configuration presented in this update demonstrates substantial improvements compared to previously evaluated models (<u>IPHC-2024-SRB025-10</u>). Driven by the goal of improved temporal generalization, the new model applies more aggressive image augmentation strategies,¹⁶ an adaptive learning rate and better tuned training parameters. These methodological enhancements contribute to improved model performance and predictive reliability.

Selected model evaluation

The selected model configuration utilized 2,799 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 (1,665) was used for training purposes. The validation set (294) 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 (30%, 840) 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 2,704 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 80 epochs. A learning rate reduction was triggered if validation loss plateaued for 40 epochs, reducing the rate by a factor of 0.6. The initial learning rate was set at 0.0002, and training was performed using a batch size of 16. A comprehensive suite of image augmentation techniques (e.g., rotation, zoom, flipping, 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 15 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 <u>Appendix C</u>.

Across ensemble runs, the model trained for an average of 288 epochs (208 effective epochs with early stopping set at 80). It achieved a normalized MSE of 0.00016 on the validation set and 0.00188 on the test set. When results were rounded to the nearest integer age, the average RMSE for the test set was 1.80. On average, the ensemble predicted the exact age correctly for 30.3% of test images, and an additional 41.7% were within ±1 year of the manually assigned age, resulting in a total agreement within 1 year for over 70% of cases.

Figure *4* illustrates the evolution of model accuracy over training epochs for one representative run. 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.

¹⁶ 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].



Figure 4. Age accuracy (measured as normalized age MSE) throughout the training process (example for seed 19).



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 statistically significant bias was observed mainly in age categories 21+ (increase from 16+ reported in <u>IPHC-2024-SRB025-10</u>). 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 2.6% of the otoliths (74 samples) used in the model were from fish aged 21 or older.





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 2.562, representing an approximate 42% increase compared to the 2019 test set. Furthermore, the proportion of predictions within ±1 year of the manually assigned age dropped by 16.7 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 8 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 8: 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 2.562 to 2.396, and the proportion of predictions within ± 1 year of the manually assigned age increased from 55.4% to 57.6%.

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 2.150.

Predicting region of collection

In September 2024, the SRB made the following recommendation:

The SRB RECOMMENDED that the Secretariat investigate using the AI to identify region of collection. Otolith shape is sometimes used as a tool for understanding mixing and stock structure and the AI may have skill in identifying region of origin (and thus mixing and migration rates) from otolith images. (<u>IPHC-2024-SRB025-R</u>, par. 47)

In response, the InceptionV3 architecture model was rewritten to perform classification task, predicting IPHC Regulatory Areas (categorical label) from otolith images. The model was trained on the 2019 FISS dataset, and performance was evaluated using three test scenarios:¹⁷

• Test set from 2019 (same year as training data):

¹⁷ Each model was run three times to account for variability due to random initialization.

The model achieved strong performance, with overall accuracy between 90% and 95%. Misclassifications were minimal and typically involved geographically adjacent areas.

(See Figure 8a: Confusion matrix – 2019 test set)

• Test set from 2024 (no fine-tuning):

When applied directly to otoliths collected in 2024, the model's predictive accuracy dropped sharply. Most images from multiple regulatory areas were misclassified as belonging to IPHC Regulatory Area 2C, suggesting a model bias toward centrally-located region.

(See Figure 8b: Confusion matrix – 2024 test set without fine-tuning)

• 2024 test set with 20% samples used for fine-tuning:

To improve temporal generalization, the model was fine-tuned using a 20% subset of the 2024 dataset, then evaluated on the remaining 80%. This approach substantially improved classification accuracy, yielding correct results for 88.4% samples. Predictions for Regulatory Areas 2B and 2C were particularly improved, with confusion concentrated around adjacent boundaries.

(See Figure 8c: Confusion matrix – 2024 test set with fine-tuning on 20% samples)

In addition, regional prediction was also evaluated using surface images (i.e., unprocessed otoliths). These models achieved promising results, with overall accuracy ranging between 87% and 91%, when trained on full sample of surface images (5,557 images). However, this evaluation was limited to data from a single year. As no multi-year dataset of surface images was available, it was not possible to assess the model's robustness or generalization across time for surface-based classification.



Panel c: 2024 test set with fine-tuning on 20% samples

Figure 9: Confusion matrices representing results from predicting IPHC Regulatory Areas (categorical label) from otolith images.

Surface images

This analysis examined whether otolith images captured prior to processing (surface images) can be used to reliably predict fish age using AI models, and how their performance compares to the use of images of processed otoliths. The goal was to evaluate both the viability and potential accuracy of surface images as a practical alternative.

Three configurations were tested:

- 1. **BB match**: The model was trained using 2,696 sectioned and baked otolith images collected during the 2024 FISS, for which matching surface images were also available (5 runs).
- 2. **Surface match**: The model was trained on the same selection of 2,696 surface images (5 runs) to allow a direct comparison under identical input conditions (sample size and age distribution).
- 3. **Surface ALL**: A model was trained using the full set of 5,557 available surface images, maximizing data size (3 runs).

The comparative analysis of otolith surface images and images of processed otoliths (see Table 1) demonstrated that surface images are a viable alternative for AI-based age prediction.

When models were trained on matched datasets, predictive performance using surface images was comparable to that of processed otoliths images, with similar test set MSE and R² values. Furthermore, the model trained on the full set of 5,557 available surface images achieved strong results, with an average test MSE of 0.00298. These findings suggest that surface images, when available in sufficient quantity, can potentially match models based on processed otoliths. This highlights the potential to streamline future otolith ageing workflows by relying on unprocessed images without compromising predictive accuracy. However, it is important to note that this evaluation was limited to data from a single year. In the absence of a multi-year surface image dataset, it was not possible to assess the temporal robustness or generalization capability of the surface-image-based models.

babba agonig.			
	BB match	Surface match	Surface ALL
Epochs trained	231	223	229
Validation MSE	0.00273	0.00298	0.00284
Test MSE	0.00315	0.00297	0.00298
R ²	0.79	0.80	0.79
Run time (VM)	159	164	345

Table 1: Average results of model configurations used to assess viability of surface images for Albased ageing.

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.

Among the models tested, the InceptionV3 architecture outperformed newer and more complex architectures such as EfficientNet and ConvNeXt. This outcome likely reflects the relatively limited size and variability of the training dataset, which favors architectures with fewer parameters and less sensitivity to overfitting. While deeper models may eventually outperform simpler ones with more data and advanced tuning, InceptionV3 currently offers the most robust and consistent performance for this application.

These results 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 (e.g., those with high standard deviation across ensemble members) can be flagged for expert review, while high-confidence outputs may be accepted directly - streamlining the ageing workflow while maintaining accuracy.

Results also showed that model performance deteriorates when predictions are made on data collected in years different from the training sample (i.e., temporal generalization is limited). However, modest fine-tuning with current-year data improved predictive performance, reducing RMSE of predictions and increasing accuracy within ±1 year of expert labels. When fine-tuning was focused specifically on uncertain samples - those with the highest variance across ensemble predictions - performance gains were even better. These findings confirm that targeted fine-tuning, guided by uncertainty, is an effective strategy for adapting models to new data while minimizing manual ageing need.

Surface images also showed promise as a practical input for ageing models. When trained on matched datasets, models using unprocessed surface images performed comparably to those using sectioned and baked otoliths. These findings point to the possibility of eliminating otolith processing steps for Al-based ageing in the future, though further multi-year evaluation is needed to confirm long-term robustness.

Despite promising progress, important limitations remain. Statistically significant bias was observed in predictions for the oldest age categories (21+), which remain underrepresented in the training dataset. Only 2.6% of otoliths used in the main model were from fish aged 21 or older, suggesting that improved model accuracy for older fish will require supplementing database in a targeted manner with images from older fish. Expanding the dataset to improve representation across all age classes especially older individuals will be essential to reduce residual bias and ensure model reliability across the full biological age range.

Finally, it is crucial to emphasize that AI-based ageing models must continue to rely on human experts, both for validation and for providing high-quality training data that reflect temporal, spatial, 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 breakand-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

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			

APPENDIX A: COUNTS OF OTOLITHS AGED BY THE IPHC

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	ВВ	14,080	16,106	1,915	21	1,500	625
2011	BB	14,451	11,391	4,592	26	1,500	676
2012	ВВ	17,896	12,902	1,639	9	1,500	1164
2013	ВВ	12,717	11,039	2,044	19	1,503	1020
2014	BB	16,194	12,606	1,476	22	1,500	1096
2015	ВВ	15,815	12,312	2,133	24	1,500	1072
2016	вв	15,113	11,618	742	21	1,502	902
2017	ВВ	12,565	10,821	1,384	15	1,500	756
2018	вв	12,935	11,013	576	39	1,499	798
2019	вв	17,716	10,711	1,640	34	1,497	925
2020	ВВ	10,323	10,568	-	34	1,413	577
2021	вв	12,253	11,051	1,444	38	1,500	547
2022	вв	9,702	10,942	1,902	39	2,334	519
2023	вв	8,506	10,932	(3,147)	(48)	(1,958)	462
2024	BB	5,770	10,474 ¹	(1,058)	(61)	(1,542)	458

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



IPHC-2025-SRB026-10

APPENDIX B: SELECTION OF MODEL RUNS

Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SETUP				**												**
Architecture	Inceptio nV3	Inceptio nV3	Inceptio nV3	Inceptio nV3	Efficient NetB4	Efficient NetB4	Efficient NetB4	Efficient NetB5	Efficient NetB5	Efficient NetB5	Efficient NetV2 S	Efficient NetV2 M	Efficient NetV2 L	ConvNe Xt	ConvNe Xt	Inceptio nV3
Max epochs	600	600	600	600	600	600	600	600	600	600	600	600	600	600	600	600
EarlyStopping patience	50	100	100	80	50	50	50	50	50	50	60	50	100	100	60	80
ReduceLROnPlateau	NA	NA	NA	40/r=0.6	NA	NA	NA	NA	NA	NA	30 /f=.8	30 /f=.8	50 / f=0.5	50 / f=0.9	30 /f=.8	40/r=0.6
Learning rate (initial)	0.0002	0.0004	0.0004	0.0002	0.0004	0.0002	0.0004	0.0004	0.0004	0.0004	0.0016	0.0004	0.0008	0.0016	0.0016	0.0002
Batch size	16	8	16	16	16	16	8	8	16	4	8	8	8	16	12	16
Image size	400	400	400	400	380	380	380	456	456	456	384	480	512	224	224	400
Dropout rate	0.2	0.2	0.2	0.2/0.25	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2/0.25
L2 parameter	0.025	0.025	0.025	.025	0.025	0.025	0.025	0.025	0.025	0.025	0.03	0.025	0.025	0.025	0.025	0.025
Augmentation ¹	NA	NA	NA	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full
RESULTS																
Validation MSE	0.00195	0.00167	0.00156	0.00170	0.00334	0.00372	0.00444	0.00414	0.00308	0.00375	0.00865	0.00223	0.00789	0.00856	0.00334	0.00163
Epochs trained	92	297	249	260	156	109	80	126	128	166	142	123	224	199	138	318
Test MSE	0.0023	0.0021	0.0020	0.0019	0.0032	0.0040	0.0044	0.0038	0.0030	0.0041	0.0087	0.0025	0.0087	0.0087	.0087	0.0019
R ²	*	*	*	.77	*	*	*	*	*	*	*	*	*	*	*	0.78
RMSE-unscaled	1.986	1.880	1.877	1.834	2.341	2.591	2.718	2.543	2.254	2.649	*	2.072	3.833	*	*	1.782
Correctly predicted	29.5%	33.6%	31.7%	31.7%	21.3%	15.6%	22.9%	31.1%	27.9%	26.9%	*	26.5%	19.3%	*	*	30.4%
Correctly predicted	75.6%	77.4%	78.8%	72.1%	55.4%	43.9%	63.9%	72.1%	75.3%	70.8%	*	75.6%	65.1%	*	*	74.4%
with ±1 year tolerance																
RUN parameters																
Machine ²	DS	DS	DS	MM	QS	QS	QS	QS	QS	VM						
Run time in hours	14.0	47.3	35.2	11	*	*	*	30.0	32.3	38.9	12.3	29.0	116.4	45.3	45	4
RESULTS for 2024																
RMSE-unscaled	2.852	2.864	2.970	2.779	3.057	3.274	*	*	*	*	*	2.801	*	*	*	2.696
Correctly predicted	18.0%	18.0%	19.3%	19.0%	17.7%	10.9%	*	*	*	*	*	15.7%	*	*	*	19.9%
Correctly predicted with ±1 year tolerance	52.5%	48.3%	50.4%	50.2%	46.4%	32.8%	*	*	*	*	*	48.9%	*	*	*	54.9%

Note: All models for randomly selected seed numbers – individual results would vary.

1: Full augmentation setup 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]. 2: Machine setups were as follows:

• QS: 11th Gen Intel(R) Core(TM) i7-11700K @ 3.60GHz; 8 cores

• DS: 12th Gen Intel(R) Core(TM) i7-12700; 12 cores

- MM: AMD Ryzen 9 5900X; 12 cores
- VM: AMD EPYC 7V12 64-Core Processor with Nvidia Tesla T4 GPU

* Indicates values not recorded for the given run.

**Indicates models selected for further investigation.

APPENDIX C: DEEP ENSEMBLE INDIVIDUAL RESULTS

Model run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE
Epochs trained	194	557	172	159	318	235	263	338	204	380	192	483	292	174	364	288
Validation MSE	0.0017	0.0015	0.0017	0.0017	0.0016	0.0017	0.0015	0.0016	0.0018	0.0015	0.0017	0.0015	0.0014	0.0016	0.0016	0.0016
Test MSE	0.0020	0.0018	0.0021	0.0022	0.0019	0.0019	0.0019	0.0018	0.0021	0.0017	0.0020	0.0017	0.0018	0.0019	0.0018	0.0019
R ²	0.776	0.797	0.756	0.749	0.783	0.784	0.779	0.794	0.764	0.804	0.774	0.809	0.797	0.785	0.796	0.783
Rum time (VM, min)	148	418	133	123	240	179	203	256	156	286	148	369	223	134	276	219
RESULTS – TEST SET																
Test RMSE unscaled	1.819	1.742	1.908	1.960	1.782	1.786	1.817	1.757	1.876	1.719	1.856	1.693	1.741	1.814	1.745	1.80
Correctly predicted	30.0%	30.6%	28.9%	23.5%	30.4%	31.3%	32.0%	31.4%	28.7%	32.5%	30.6%	32.1%	33.6%	29.0%	30.4%	30.3%
Correctly predicted with ±1	72.0%	74.5%	69.8%	64.6%	74.3%	71.3%	73.3%	74.4%	69.5%	74.5%	69.2%	75.1%	72.6%	71.3%	74.2%	72.0%
year tolerance																
RESULTS – 2024 IMAGES																
RMSE	2.509	2.472	2.598	2.844	2.514	2.539	2.631	2.498	2.613	2.477	2.660	2.548	2.481	2.519	2.518	2.562
Correctly predicted with ±1	56.8%	57.4%	55.4%	52.7%	55.9%	55.1%	55.2%	55.5%	54.0%	58.8%	52.1%	57.1%	56.3%	52.1%	56.0%	55.4%
year tolerance																
RMSE – fine-tuned on 20%	2.378	2.350	2.451	2.418	2.328	2.404	2.396	2.389	2.440	2.331	2.493	2.379	2.408	2.444	2.334	2.396
images																
Correctly predicted with ±1	59.7%	58.0%	54.4%	56.2%	59.1%	56.5%	58.0%	57.5%	57.0%	59.7%	56.3%	58.8%	57.0%	57.1%	58.4%	57.6%
year tolerance- fine-tuned on																
20% images																
RMSE – fine-tuned on 20%	2.151	2.105	2.142	2.211	2.069	2.133	2.159	2.108	2.270	2.073	2.280	2.084	2.116	2.260	2.089	2.150
images with highest standard																
deviation	50.00/	50.40/	50 70/	50 70/	00.00/	50.00/	57.00/	50.00/	50.404	57.00/	54.004	00.50/	50.404	50.00/	00.00/	
Correctly predicted with ±1	56.3%	59.4%	58.7%	53.7%	60.9%	59.0%	57.6%	59.3%	52.1%	57.9%	51.6%	60.5%	59.1%	52.8%	60.2%	57.3%
year tolerance - The-tuned on																
20% images with highest																
Stanuaru uzvialion		1	1	1	1	1	1	1	1	1	1	1	1	1	1	