INTERNATIONAL PACIFIC

HALIBUT COMMISSION

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

Agenda item 4.2.3:
IPHC-2025-SRB026-10

(B. Hutniczak, J. Forsberg, K. Sawyer Van Vleck & K. Magrane)

Purpose

- To summarize the information available on the use of artificial intelligence (AI) for determining the age of fish from images of collected otoliths
- To provide an update on the exploratory work of implementing an AI-based age determination model for Pacific halibut

Why AI-based model?

- 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.

Key progress updates

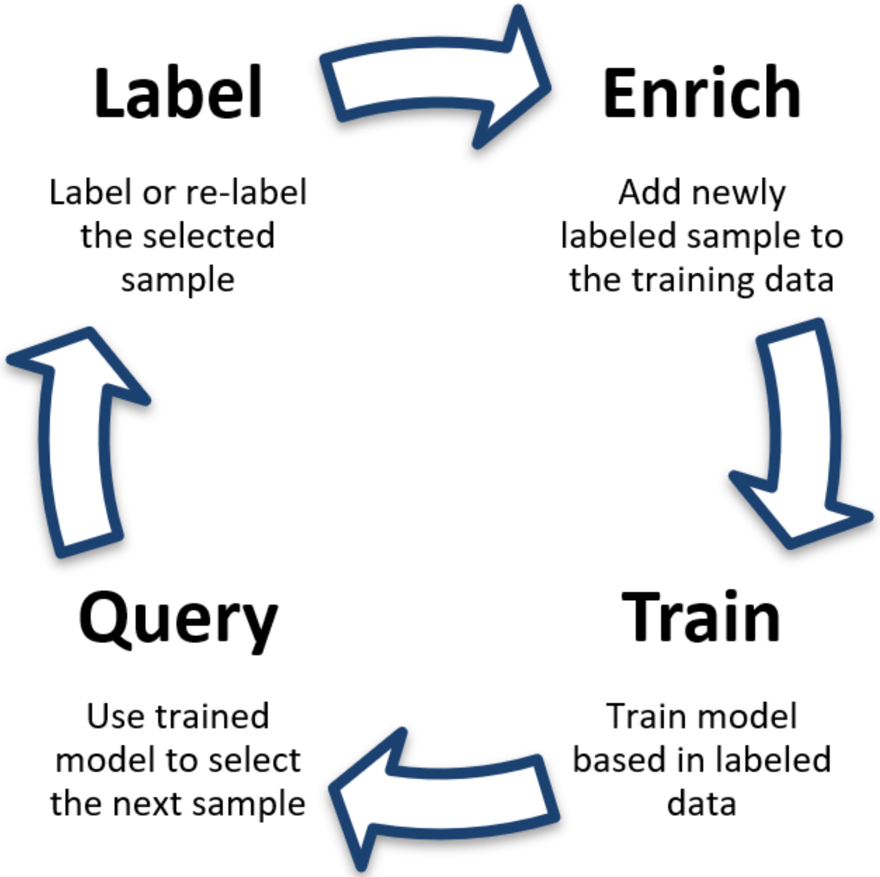
- **Testing various deep learning architectures** to identify the optimal approach given the available otolith images
- **Evaluating model temporal generalization** by comparing age predictions from a model trained on images from one year to those from a different year
- Assessing differences in model 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**

Model framework

The proposed approach integrates AI-based age determination and traditional ageing methods for maximum accuracy of the estimates.

Climate-readiness:
Training the model with inputs that capture temporal changes

- ➢ increasingly important in the face of changing environmental conditions and climate change



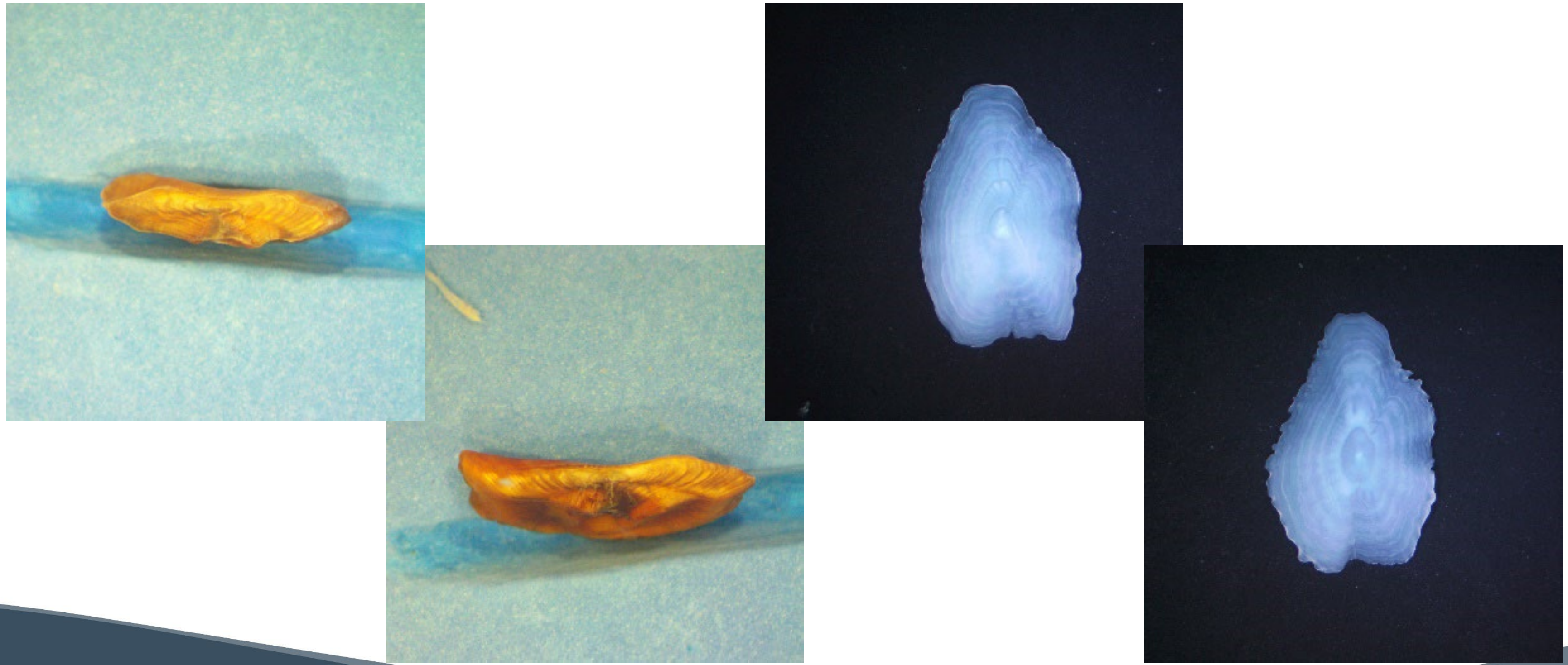
Modeling approach

- Application of **convolutional neural network (CNN) model**, a type of deep learning approach
  - In CNNs, the layers are structured as stacks of filters, each recognizing increasingly abstract features in the data.
- Application of **image regression** - predicting a continuous variable from an image
  - Pacific halibut is a long-living species - oldest Pacific halibut on record were aged at 55 years
- Implementation - TensorFlow and Keras libraries, repurposing **Inception V3** model from Google:
  - Input → InceptionV3 (feature extractor) → Regressor → Output
  - Alternatives tested: EfficientNet variants (EfficientNetB4, EfficientNetB5, EfficientNetV2 S/M/L) and ConvNeXt

Database

- Since 1925, over 1.5 million otoliths have been aged and stored for potential future use. ← **<u>unique resource for AI training</u>**
- Aged otoliths are sectioned (broken in half) and baked to enhance the contrast between the growth rings, then photographed.
- In addition, IPHC is in the process of creating complementary database comprising labelled images of otoliths captured prior to processing.

- Pictures are taken with AmScope 8.5MP eyepiece cameras



Comparison of model architectures

- Several modern CNN architectures were systematically evaluated to identify the most suitable approach for ageing Pacific halibut using otolith images. Models tested included included InceptionV3, EfficientNet (B4, B5, V2 S/M/L) and ConvNeXt.
- 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**.
- This suggests that the more complex models struggled to extract meaningful age-related features, <u>likely due to a combination of limited training data and overfitting driven by model complexity</u>.
- **Current approach:** given current dataset limitations, <u>the simpler structure of InceptionV3 provides more reliable and robust predictions</u>.

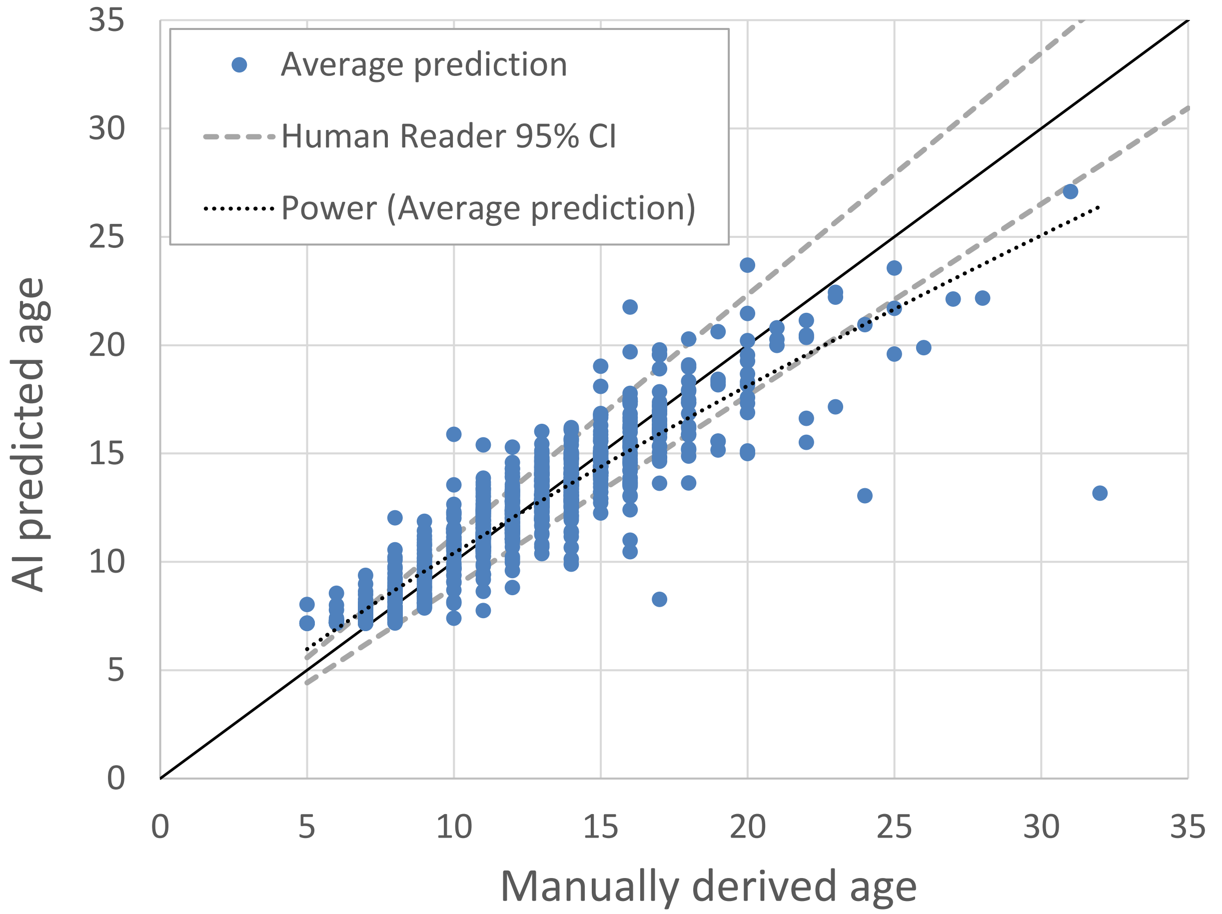
Latest runs results (InceptionV3)

- Main setup (2019 FISS): 2,799 images (1,665 train, 294 validation, 840 test)
  - Test set is used to assess the performance of the model after training, providing an unbiased evaluation of its generalization capability to new, unseen data.
- Testing the temporal generalization (2024 FISS): 2,704 images
- Surface images (2024 FISS): 5,557 images (2,696 matched with BB images)

- All images resized to 400x400 pixels; broken otoliths excluded
- Epochs: 600; patience: 80
- Learning rate: initial at 0.0002 with reduction triggered if validation loss plateaued for 40 epochs
- Batch size: 16
- Deep ensemble approach: the model was trained 15 times, each with a different random seed
- The model trained for an average of 288 epochs. It achieved a normalized MSE of 0.0016 on the validation set and 0.0019 on the test set

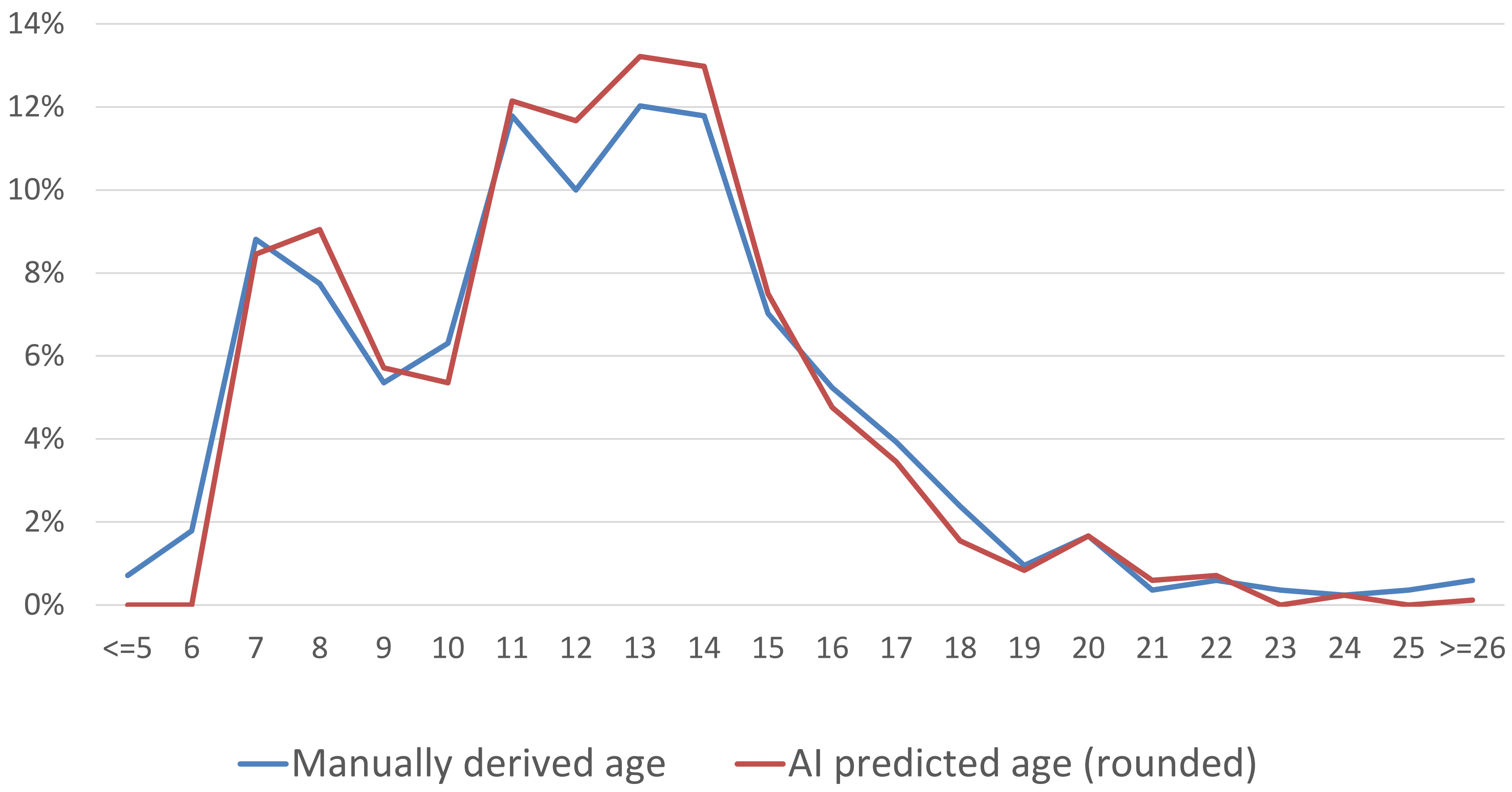
Base run results (1/3)

The model achieved average RMSE in the test set of **1.80** when calculated for rounded results. On average, the ensemble **correctly predicted age for 30.3% individuals**, with an **additional 41.7% being within 1 year of error**.



Comparison between manually derived age with AI predicted age.

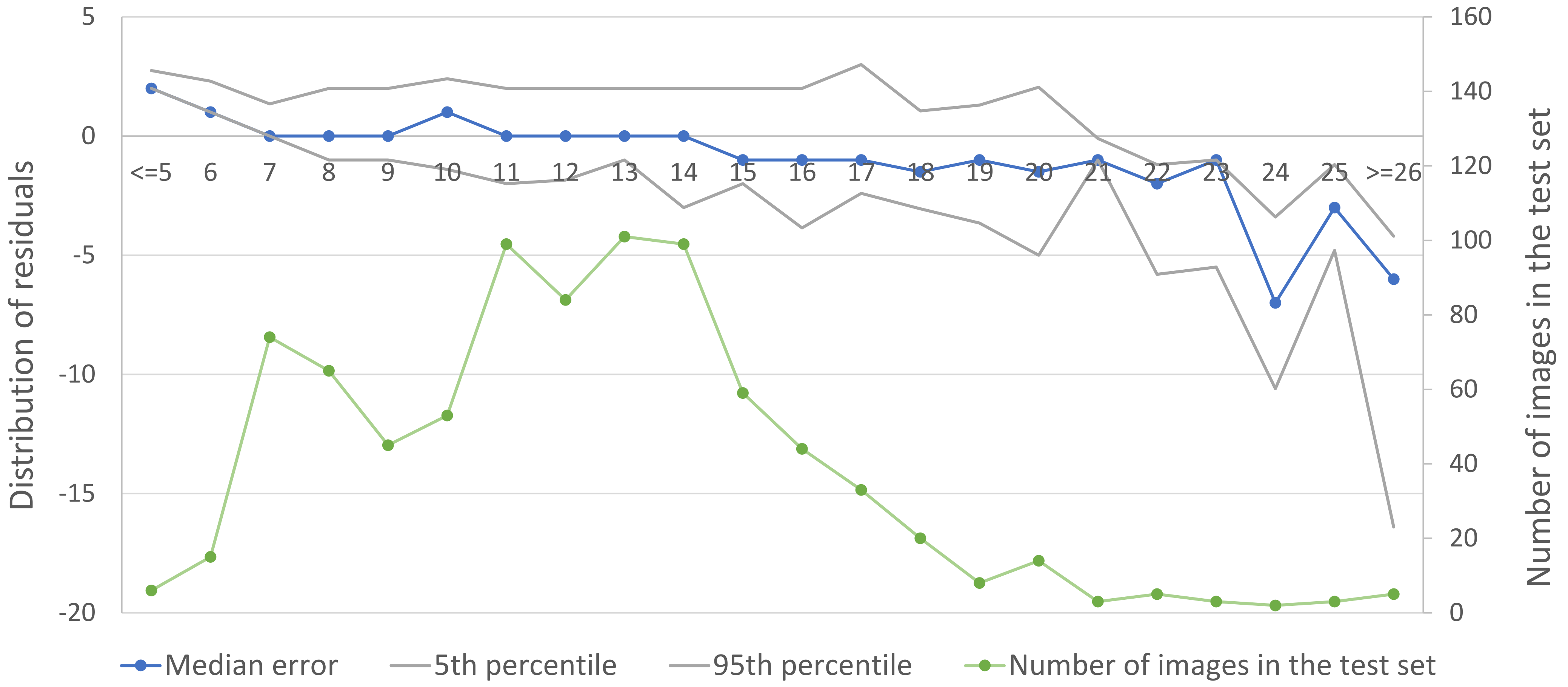
Base run results (2/3)



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

Base run results (3/3)

Statistically significant bias was observed in age categories 21+ (previously 16+), where the number of observations remains low despite an overall increase in sample size.



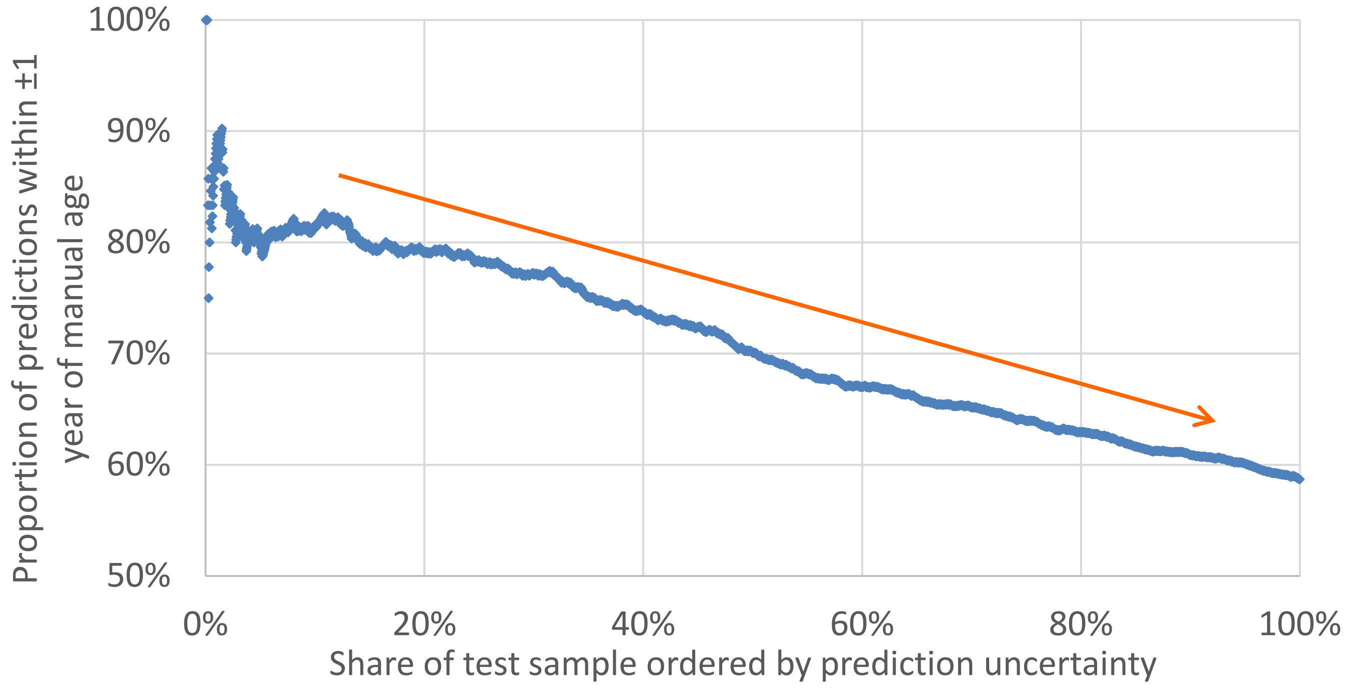
Distribution on residuals and number of images by age in the test set.

Temporal generalization

- A model trained on 2019 FISS otolith images was evaluated on 2024 data to assess temporal generalization.
- RMSE increased to 2.56 (42% increase compared to within-year predictions).
- The proportion of predictions within ±1 year of the manually assigned age decreased by 16.7 percentage points.

Fine tuning applied:

- 20% randomly selected images – RMSE = 2.40
- 20% images with the highest standard deviation across ensemble predictions – RMSE = 2.15



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).

Predicting region of collection (1/2)

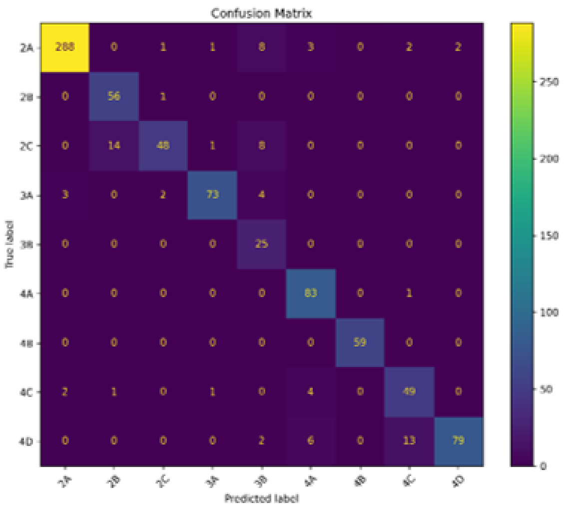
- The InceptionV3 architecture model was rewritten to perform classification task, predicting IPHC Regulatory Areas from otolith images. Model was trained on 2019 FISS otolith images.
- Performance evaluated for:
    a. Test set from 2019 (same year as training data)
    b. Test set from 2024 (no fine-tuning)
    c. Test set from 2024 with 20% samples used for fine-tuning

Predicting region of collection (2/2)

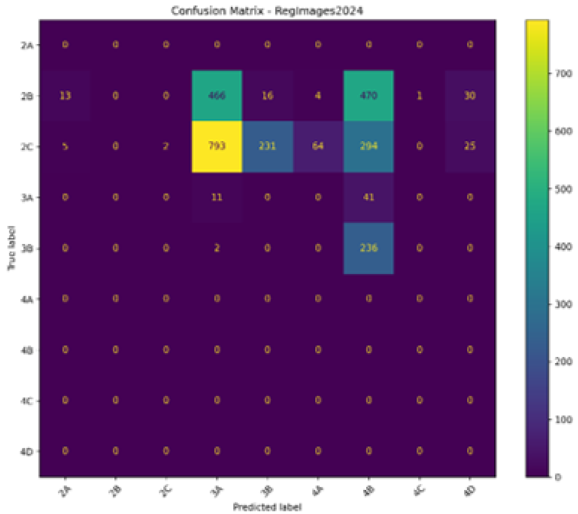
a. Within-year accuracy was between 90% and 95%. For surface images, between 87% and 91%.

b. Poor performance, most images misclassified as belonging to IPHC Regulatory Area 3A, suggesting bias toward centrally-located region.

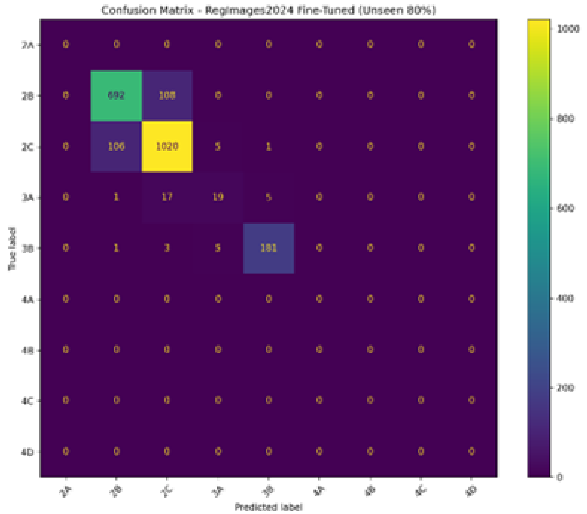
c. Fine-tuning improved generalization – accuracy for the remaining 80% of images increased, yielding correct results for 88% samples.



Panel a: 2019 test set



Panel b: 2024 test set without fine-tuning



X axis: Predicted label

Y axis: True label

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

Surface images analysis (1/2)

This analysis examined whether otolith images captured prior to processing (surface images) can be used to reliably predict fish age, and how model performance compares to the use of images of processed otoliths.

Three configurations were compared:

a. **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).

b. **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).

c. **Surface ALL**: A model was trained using the full set of 5,557 available surface images, maximizing data size (3 runs).

Surface images analysis (2/2)

	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 (min/VM)	159	164	345

*Evaluation limited to data from a single year

Conclusions

- The ongoing advancement of AI technologies in the field of marine science offers considerable **potential to improve the efficiency of age determination** of Pacific halibut using otolith images.
- Preliminary results presented here suggest that AI could serve as a **promising alternative to the current ageing protocol**, which relies entirely on manual age reading.
    - Performance may be further improved by expanding the training set, particularly with more images from older age categories.
    - A smaller dataset currently favors the use of simpler model architectures.
- **NEW**: **practical value of the deep ensemble framework** - per-sample uncertainty estimates help identify potentially unreliable predictions, allowing for targeted expert review.
- **NEW**: **targeted fine-tuning** offers an effective strategy for adapting models to new data, including better temporal generalization.
- An **adaptive approach** will continue to partially depend on trained readers for **capturing temporal changes**.
- **AI is evolving rapidly**, and adapting to new developments may further improve results over time.

INTERNATIONAL PACIFIC

HALIBUT COMMISSION

https://www.iphc.int/