INTERNATIONAL PACIFIC

HALIBUT COMMISSION

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

Agenda item 4.2.3:
IPHC-2026-SRB029-INF02
(B. Hutniczak)

Purpose

- To summarize the current knowledge on the use of artificial intelligence (AI) for determining the age of fish from images of collected otoliths
- To provide 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.

Key progress updates

- **[SRB request]** Use case comparison – comparison of break and bake (B&B) and CNN
- **Mix method approach** - proposed operational design integrating manual and AI-based ageing into a single production workflow
- Estimation of **ageing error for 50/50 BB + AI mix**
- Evaluation of new model architectures – **ConvNeXtSmall @ 480x480** outperforms InceptionV3 presented at SRB028
- Assessment of performance of using **Z-stack images**

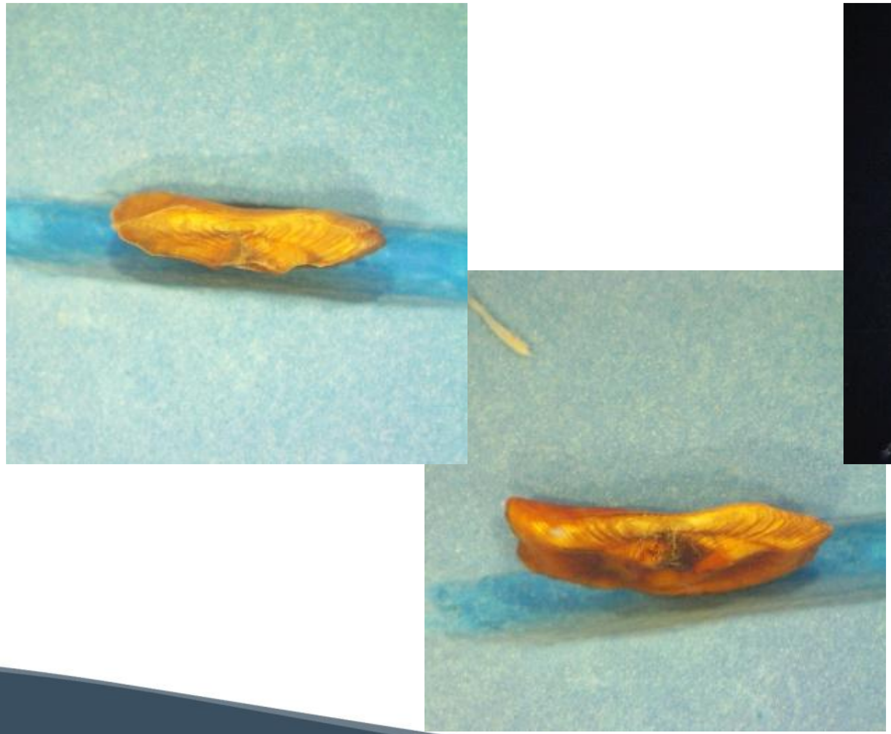
Modeling approach

- 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, with **ConvNeXtSmall at 480×480 resolution** as preferable configuration
  - earlier testing found alternative architectures underperforming relative to InceptionV3 - with the subsequent expansion of the image database, this conclusion no longer holds

Database

- Since 1925, over 1.6 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.
- Complementary set of otolith images is captured prior to processing (surface images).

- Pictures are taken with AmScope 8.5MP eyepiece cameras



Use case comparison (summary)

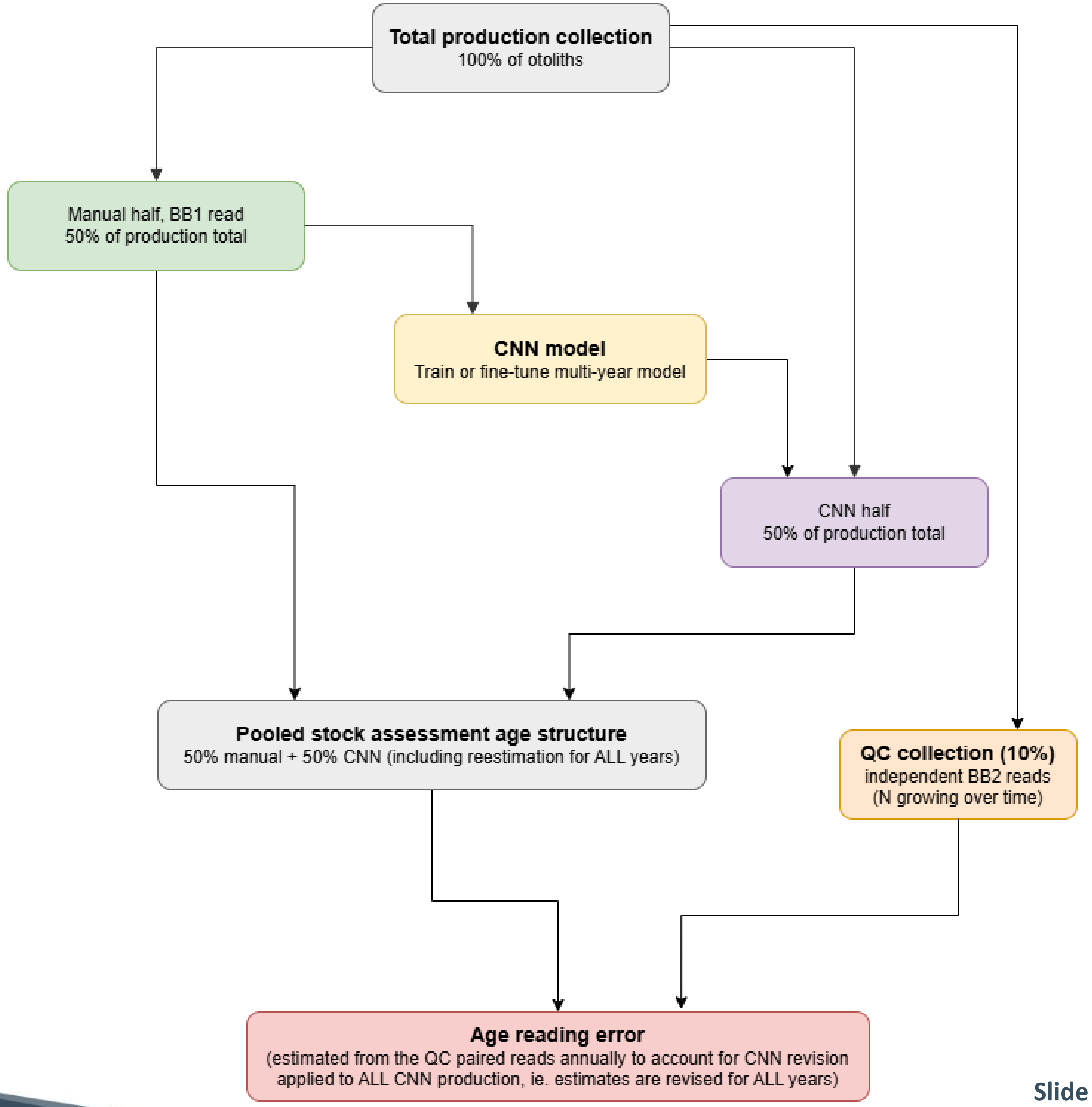
	Bake and break	AI-based ageing (CNN)
Use cases	Reference ageing method for stock assessment Source of validated ages used to train AI model	Supplement to manual ageing Could be used in mix method approach (see 50/50 mix method approach proposed design)
Development costs	None (established method) Real cost is training a new ager - considerable, months to proficiency	Developed in-house - intermittent staff time (decreasing with AI-assisted coding) Virtual machine cost (model testing/selection)
Production costs	Ageing ~$3 to $6+ per otolith - excludes training; higher as ager gains proficiency	Imaging ~$0.40-50 per otolith - requires ~15-min training Compute ~$300 per model build
Benefits	Established method accepted for assessment; No new infrastructure	Low marginal cost Readily scalable Ages can be rerun on the same images with better models over time
Timelines and decision points	N/A (established) Ongoing reader precision monitoring	50/50 BB/AI mix showed ageing error close to BB-only Temporal generalization not fully proven – but 50/50 mix method offers built-in adaptation to change

Mix method approach

Proposed operational design integrates manual and AI-based ageing into a single production workflow

Benefits:

- Easier to scale
- CNN updated annually
- Training data accumulates
- Ageing error is quantified from a quality-control (QC) subset



Ageing error of a 50/50 BB + AI mix

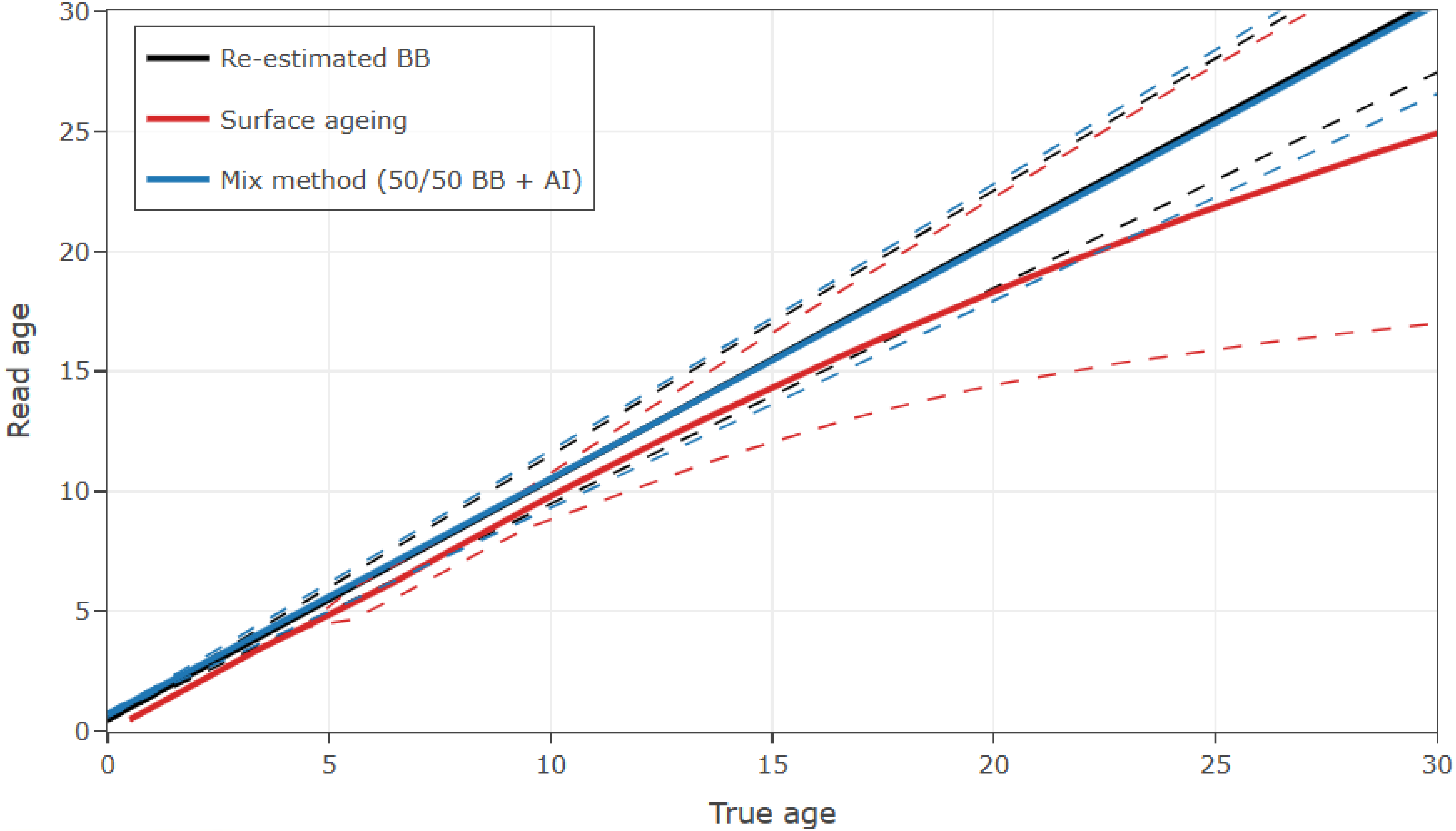
Ageing error was predicted for a simulated 50/50 random mix of break-and-bake (BB) reads and AI (CNN) predictions (Punt et al., 2008)

- This approach fits paired age readings of the same otoliths to jointly estimate bias and imprecision, with one reading method assumed to be unbiased

The results indicate that:

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

Ageing error results



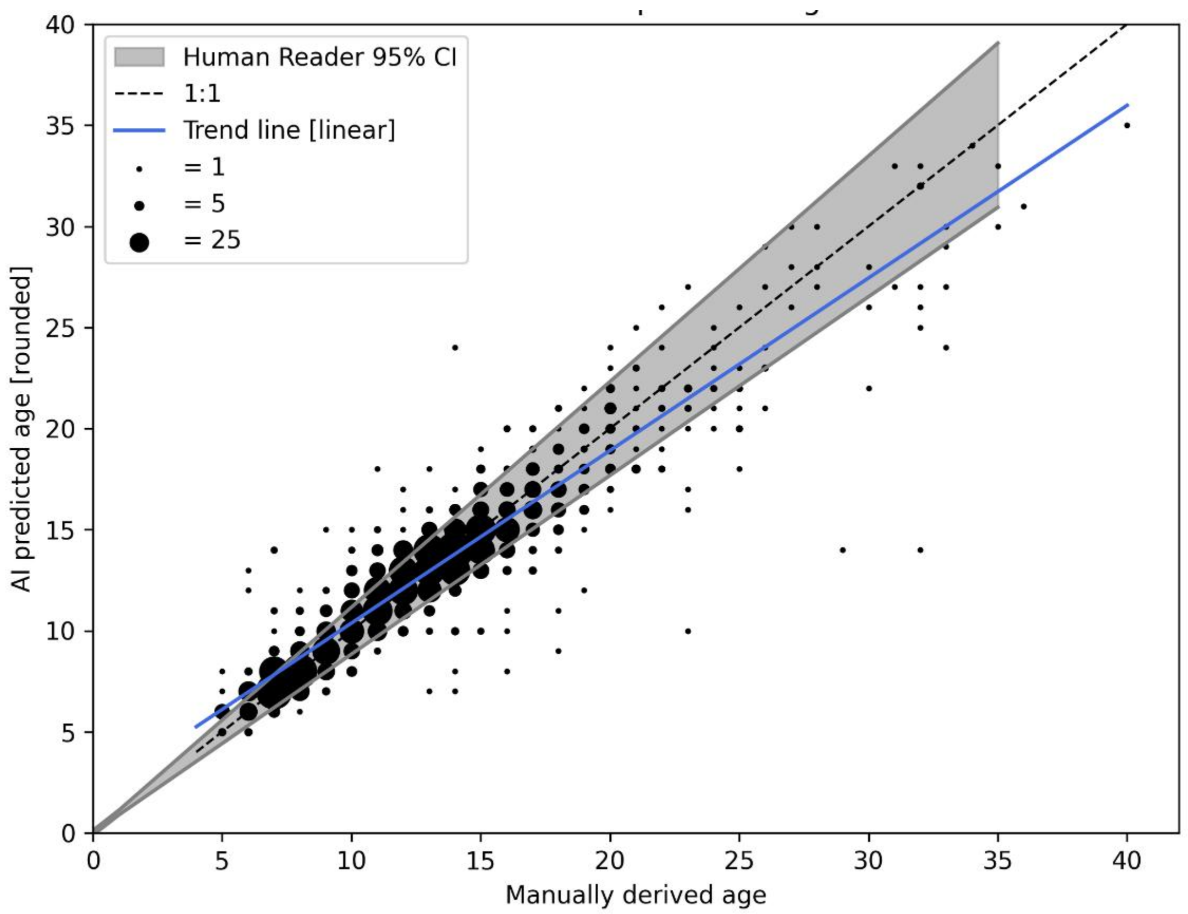
Selected run results

Same images – new architecture

- 11,107 images (7,740 train, 1,367 validation, 2,000 test) – ConvNeXtSmall @ 480 pixels, 7-seed ensemble

The model achieved average RMSE in the test set of **1.62** (↘ from 1.68) when calculated for rounded results. On average, the ensemble **correctly predicted age for 36.7% individuals** (↗ from 35.5%), with an **additional 43.5% being within 1 year of error** (↗ from 41.0%).

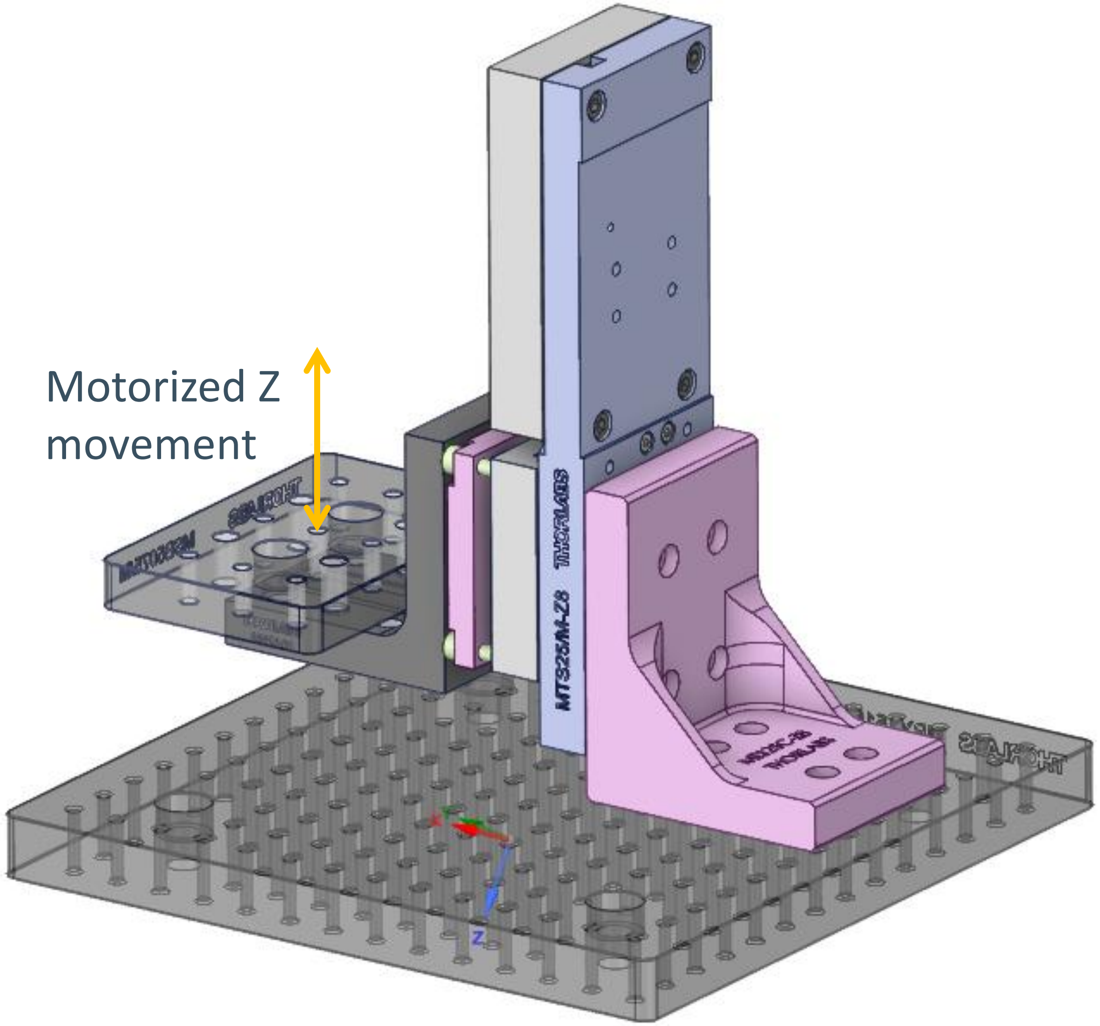
➢ **Total agreement within 1 year for 80.2% of cases**



Comparison between manually derived age with AI predicted age.

Z-stack imaging experiment

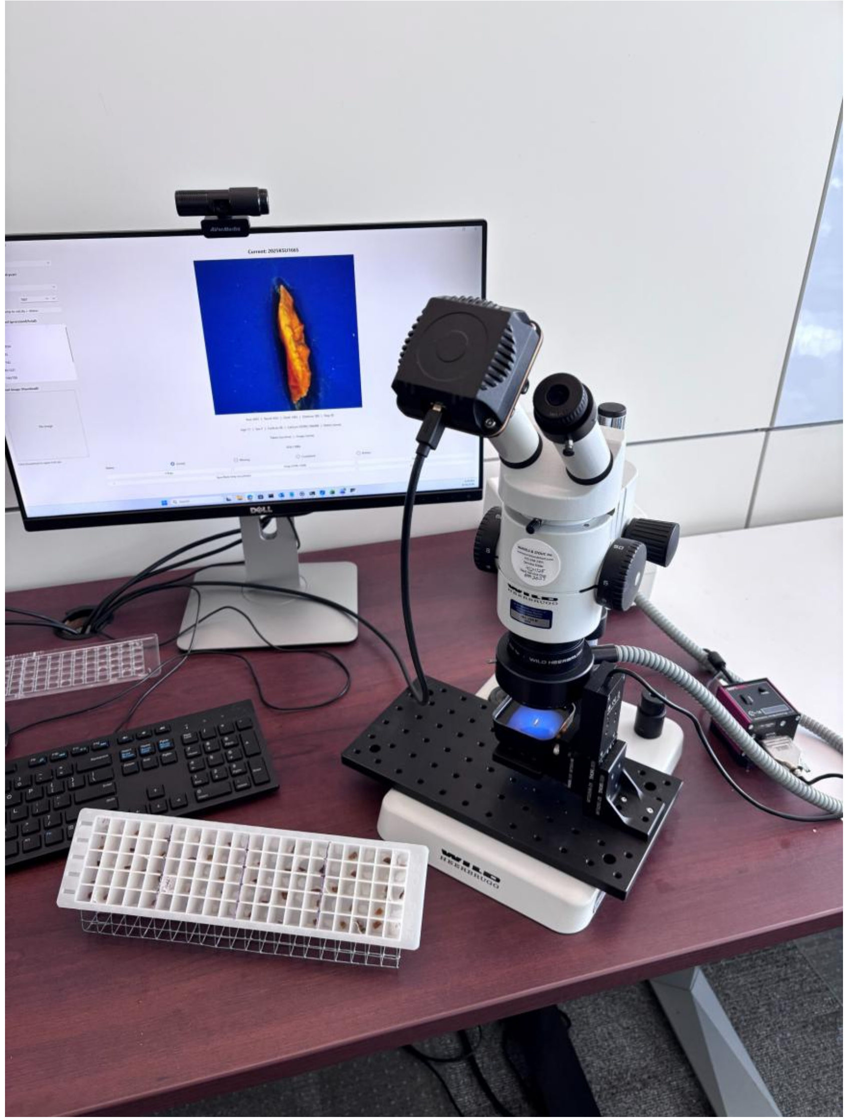
Z-stacking is an imaging technique in which multiple images are captured at different focal planes along the Z-axis and then combined to produce a single image with extended depth of field.



<u>**GOAL:**</u> Sharper, fully focused otolith images.

Z-stacking results

- **Design:** rigorous paired comparison - variants trained and tested on identical specimens, setups differed only in image input (full stack vs. single sharpest slice vs. conventional image); 1,057 paired specimens (152 held out for test)
- **Result: no accuracy gain.** No variant beat the single sharpest slice; differences averaged zero (+0.0, −0.7, +0.7, −2.0, +2.0 pp) and none exceeded the ~3 pp variation from training randomness alone
- **Depth signal is real but unused:** the in-focus region genuinely shifts with depth (88% of stacks), yet the model extracted nothing beyond what the single image already provided
- **Still useful operationally:** Z-stack removes the need for a person to choose a focal plane, making focus selection automatic and repeatable and enabling **unattended batch imaging**
- **Caveats:** pilot scale (1,057 vs. 11,107 images) at 384 pixels input; whether depth helps at higher resolution is untested



Conclusions

- **CNN ensemble is now near human-reader agreement**: 80.2% within ±1 year, 74.4% within the manual-reading confidence interval (**ConvNeXtSmall**, 480x480 pixels)
- **Architecture matters**: newer backbones (ConvNeXt, EfficientNetV2M) now outperform the earlier InceptionV3 model, enabled by the larger dataset
- **Proposed 50/50 BB + AI mixed method**: same total ageing effort (~110%), but specialist manual burden roughly halved (~110% → ~60%)
- **Ageing error of the mix is close to break-and-bake**: slightly less precise and slightly biased, but substantially closer to BB than to surface ageing - acceptable for assessment on current data (Punt et al., 2008)
- **Built-in adaptation to change**: 50% fresh expert reads every year adapt the model to current-year data, capturing inter-annual trends rather than relying on a static model
- **Supporting findings**:
  - surface imaging remains worse than BB (BB processing still required);
  - Z-stack imaging gave no accuracy gain in a pilot
- **Next steps**: extend training and testing across multiple years/decades to confirm multi-year reliability; human expertise remains essential for validation and oversight

INTERNATIONAL PACIFIC
HALIBUT COMMISSION

https://www.iphc.int/