Fisheries New Zealand

Tini a Tangaroa

Development of deep learning approaches for automating age estimation of hoki and snapper

New Zealand Fisheries Assessment Report 2021/69

B.R. Moore,
Z.T. A'mar,
A.C.G. Schimel,
C. Ó Maolagáin,
S.D. Hoyle

ISSN 1179-5352 (online)
ISBN 978-1-99-101960-8 (online)

November 2021

New Zealand Government

Requests for further copies should be directed to:

Publications Logistics Officer
Ministry for Primary Industries
PO Box 2526
WELLINGTON 6140

Email: brand@mpi.govt.nz
Telephone: 0800 00 83 33
Facsimile: 04-894 0300

This publication is also available on the Ministry for Primary Industries websites at:
http://www.mpi.govt.nz/news-and-resources/publications
http://fs.fish.govt.nz go to Document library/Research reports

© Crown Copyright – Fisheries New Zealand

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Moore, B.R.[1]; A'mar, Z.T.; Schimel, A.C.G.; Ó Maolagáin, C.; Hoyle, S.D. (2021). Development of deep learning approaches for automating age estimation of hoki and snapper.

***New Zealand Fisheries Assessment Report 2021/69.* 33 p.**

The ages of fish are a key input to fisheries assessment models. However, preparing and reading otoliths can be expensive and time-consuming, and age interpretation can be subjective and uncertain. Recent trials conducted in New Zealand on hoki (*Macruronus novaezelandiae*) and snapper (*Chrysophrys auratus*) indicate that advances in machine learning may make it possible to improve the efficiency of ageing, with potential to reduce both biases and long-term costs. The current project builds upon work conducted in recent trials, by 1) improving the consistency and increasing the number of images available to the models and 2) undertaking further model development using the captured images.

To ensure that model development and associated outputs were relevant to current ageing practices, imaging focused on bake-and-embed prepared sections for hoki and the whole sister otolith, and break-and-burn prepared otoliths for snapper. All samples used in this study had been aged previously by human readers. A random-stratified approach was used to select samples, to ensure that developed algorithms were robust to potential effects of sex or collection location on age interpretation. This resulted in 1068 individual hoki and 520 snapper being selected for imaging and age estimation.

For hoki, ten image types were captured, including three images of the whole otolith and seven images of the bake-and-embed prepared otolith, at various orientations and magnifications. For snapper, six image types were taken of each sample, at various magnifications under reflected or ultraviolet light. Prior to inclusion in the age estimation algorithms, captured images were modified by a variety of processes, including image segmentation (i.e., removal of backgrounds), resizing, and binarisation.

A convolutional neural network (CNN) designed for object recognition was adapted to estimate age using the captured otolith images. For each species, the model was trained on a subset of images (~80% of the total number of images), validated against a smaller subset (~10%), and tested against a third subset (containing the final ~10 % of images). Models were run using unmodified and modified images.

Overall, models run on hoki otolith images outperformed those run for snapper. For hoki, mean squared error (MSE) values of models run on unmodified images ranged from 3.91 to 2.33, whereas correct agreement with human readers ranged from 25.2% to 35.5% for test data subsets. For snapper, MSE values for models run on unmodified images ranged ~9.1–35.2, with correct agreement with human readers ranging from 13.2% to 37.7%. For both species, models run on low magnification images in which backgrounds had been removed generally outperformed those using on unmodified images. Models run on individual image types outperformed those using multiple image types for each sample.

There is potential for CNN models to derive age estimates from otolith images, although further development and testing is required before such an approach could be used to conduct routine ageing. The key next steps towards the implementation of the technique for routine age estimation include:

1. evaluate the hoki model(s) against the current otolith reference collection;
2. improve / automate image capture for quality and efficiency;
3. further develop the image segmentation model;
4. resolve issues around resizing of the images within the CNN model;
5. improve understanding of the features the models are trained on; and
6. further develop an integrated approach using multiple image types and other data inputs.

[1] All authors: National Institute of Water and Atmospheric Research (NIWA), New Zealand.

1. INTRODUCTION

Reliable estimation of the ages of fish is integral to fisheries management. It is a key requirement for numerous components of age-based stock assessments, including for estimating growth, age at recruitment and sexual maturity, longevity, mortality rates, population age structure, and age-dependent fishing gear selectivity. Globally, it has been estimated that well over one million fish are examined each year to estimate age (Campana & Thorrold 2001). For most bony fishes, age is estimated by enumerating periodically-deposited growth marks in calcified structures, in particular scales, bones, fin rays, and otoliths (Francis et al. 1992, Welch et al. 1993, Horn & Sullivan 1996, Campana & Thorrold 2001, Zhu et al. 2015), with otoliths the most commonly-used structure, particularly in recent years.

Preparing and reading otoliths, or other hard parts, for age estimation can be expensive and time-consuming. Recent estimates suggest that reading accounts for approximately half of the costs involved with age estimation from extracted otoliths for inshore fish species in New Zealand, with preparation accounting for the remaining half (Jeremy McKenzie, NIWA, pers. comm.). Moreover, age estimation can be inherently subjective and uncertain. Individual readers may interpret the same otolith differently, and an individual reader's interpretation can change over time. Differences between readers, or within readers over time, can result in long-term changes in interpretation which has the potential to bias age estimations and stock assessments. Improving the consistency of ageing through automated age determination could increase the replicability of age estimation and improve the reliability of management advice.

Machine learning-based methods of fish age estimation have been explored for many years. Early studies employed artificial neural networks, which are computational structures consisting of units referred to as neurons, organised in layers. Results from these studies generally report age estimates that are less precise than those obtained from experienced otolith readers, particularly for younger and older components of tested samples (Robertson & Morison 1999, Fablet & Le Josse 2005).

In the last few years, however, there has been considerable progress in machine learning due to improved algorithms, greater computing power, and wider availability of digital training data. Central to improving the algorithms has been increasing the number of layers in neural networks, a process often referred to as deep learning, and the development of convolutional neural networks (CNNs). These are unlike previous deep learning neural networks, in which the lower layers learn to distinguish between primitive features (e.g., sharp edges or colour transitions), and subsequent layers then learn to recognise more abstract features. In a CNN, the layers are organised as a stack of convolutions, applying the same filters across the whole image. A key advantage of this process this that it greatly reduces the number of parameters to be learned, which in turn reduces the amount of data and computation necessary for training (Abadi et al. 2015, Moen et al. 2018).

With these advances, many new applications have become possible, including the potential to automate otolith age estimation. In a preliminary study, Moore et al. (2019) investigated the feasibility of using a CNN approach to estimate ages of New Zealand snapper (*Chrysophrys auratus*) and hoki (*Macruronus novaezelandiae*) from otolith images. For each species, the model was trained on a collection of images of fish previously aged by human readers (n = 687 and 882 for snapper and hoki, respectively). After training, the model gave the same age as the human reader for 47% of snapper in a test dataset, with a further 35% of ages estimated within ±1 year of the human reader estimate of age. For hoki, the model gave the same age as the human reader for 41% of individuals.

These promising results were achieved despite the application of minimal image or model optimisation. Standardising and optimising the input images, as well as optimising the ageing algorithms used, may significantly improve the ability to estimate age. The current project aimed to build on the results of Moore et al. (2019) by improving the quality of images available for age estimation by machine learning and refining the ageing algorithm used. The objectives of this project (SAM2019-02) were as follows:

1. to develop a reference library of high-quality, standardised otolith images for hoki and snapper for use in developing an automated ageing system using machine learning;

2. to use these images and associated data to train and optimise a reliable ageing algorithm for hoki and snapper.

2. METHODS

2.1 Fish selection

Considerable variation in otolith morphology can occur between stocks and sexes for a given fish species (e.g., Bose et al. 2017, Parmentier et al. 2018). Accordingly, a stratified approach to selection of samples was used in this study, with the aim of balancing sample numbers across stock (management area for snapper), sex, and age (from human reader estimates). All samples were selected from the *age* database (Mackay & George 2017), on the basis of the below criteria and in relation to the age estimate provided in the *agreed_age* field from the *t_age* table.

2.1.1 Hoki

All hoki examined in this study had been processed for ageing using bake-and embed methodologies. For hoki, approximately equal numbers of otoliths were selected from fish from the eastern and western stocks. For the eastern stock, samples were obtained from fish collected predominantly from the Cook Strait spawning ground during 2012–2018. A preliminary query of the age database revealed insufficient numbers of old fish (i.e., ≥ 15 years) were available over this period. To increase the sample sizes of older fish, additional samples were selected from this area, as well as the Central East, and Chatham Rise fishing areas, from collections dating back to 1998.

For the western stock, samples were obtained from fish collected predominantly from four areas: the west coast South Island, WC25, Puysegur, and Challenger fishing areas, with samples from the latter area restricted to those fish caught south of Cape Foulwind, to improve the likelihood that they represented the western stock. Samples were selected from fish caught during 2012–2018, although additional samples of older fish (i.e., > 13 years) were taken from collections from these four areas dating back to 1998. To further increase the sample sizes of older fish for the western stock, additional samples were taken from the southern South Island, Stewart Island, and the Sub-Antarctic islands fishing regions (encompassing the Auckland Islands, Campbell Plateau (including Campbell Island), Pukaki Rise, Snares Islands, Southland, and Stewart Island fishing areas).

From the available samples in each area, approximately equal numbers of males and females were initially selected within each 1-year age class, to a maximum of 15 individuals per age class. Fish within each age class were selected randomly, with no prior consideration of the characteristics of their otoliths. However, it was evident that not all selected individuals had otoliths suitable for imaging (e.g., the whole otolith of the pair was broken). These were substituted by individuals in the same age classes of the same sex, or, if no suitable replacement was available, by an individual in the same age class of the opposite sex, or an individual of an adjacent age class of the same sex. This resulted in a small number of age bins with > 15 individuals (Figure 1). A total of 554 and 514 individuals were selected for imaging from the western and eastern stocks, respectively, resulting in 1068 hoki being selected.

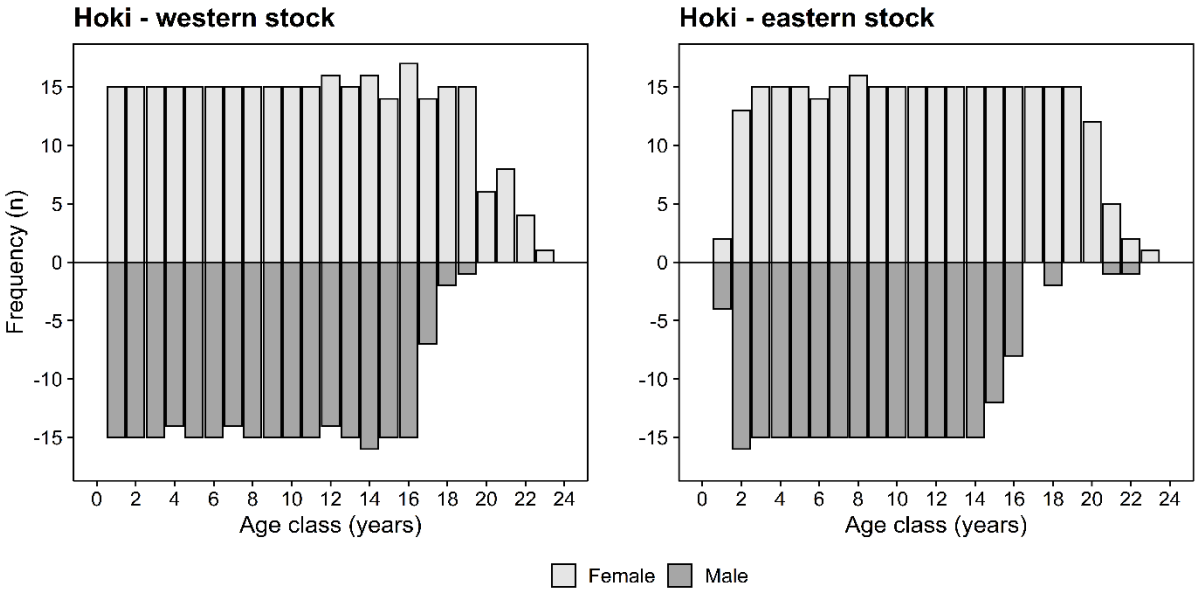


Figure 1: Age structure of hoki for which otolith images were captured. Left column = western stock, right column = eastern stock.

2.1.2 Snapper

For snapper, imaging focused on break-and-burn prepared otoliths. All samples used in this study had been aged previously by human readers and were selected from the *age* database. To ensure any stock-related geographical variation in otolith morphology was accounted for, samples were taken from the main fishery management areas where the majority of fish have been historically caught (and thus where samples had previously been collected), namely SNA 1, SNA 2, SNA 7, and SNA 8. An even spread of samples across ages and age groups was desired, however this was not possible for the older age classes due to low numbers of aged fish (Figure 2). An approximate 1:1 ratio of females:males was targeted; however this was not possible in every age group due to insufficient numbers of one sex or the other.

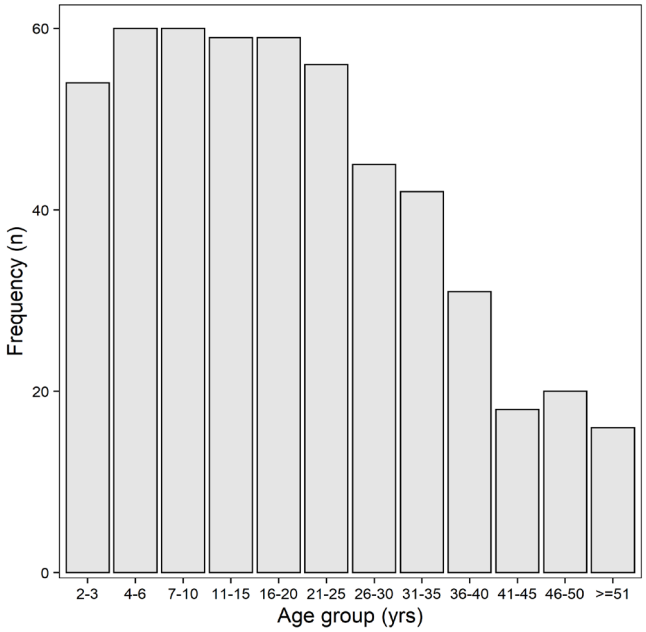


Figure 2: Age structure of snapper for which otolith images were captured.

2.2 Image capture

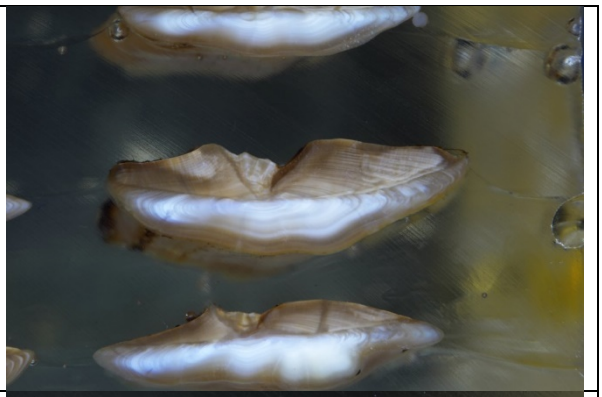
Images were captured with a Nikon DS-Ri2 camera attached to a Nikon SMZ25 stereomicroscope. Illumination was provided by a Schott KL2500 LED light source using a fiber-optic ring light to provide even lighting for sectioned embedded blocks of hoki otoliths. The same light source was used for snapper using adjustable gooseneck arms, orientated at ~45° to the burnt otolith face. All images were captured at a maximum resolution of 4908 x 3264 pixels per image with a standard exposure time of 600 ms.

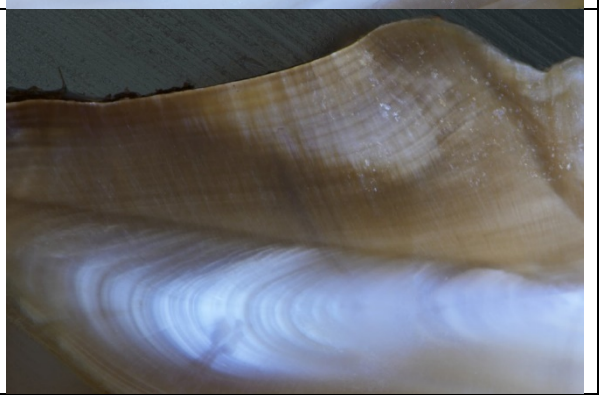
2.2.1 Hoki

Ten image types were collected for hoki. These included three types for whole otoliths, and seven types for bake-and-embed sectioned otoliths (Table 1). Autowhite balance was set at RED=4.71 and BLUE=1.55 for all image captures.

Table 1: Examples of images taken for use in age estimation models for hoki (continued on next two pages).

Image number	Description	Example
1	Whole otolith, lateral face, 4x magnification, under water with reflected light	
2	Whole otolith, medial face, 4x magnification, under water with reflected light	
3	Whole otolith, lateral face, 4x magnification, under water with transmitted light	

4	Baked otolith, whole cut face, 11x magnification under reflected light	
5	Baked otolith, ventral arm, 30x magnification under reflected light	
6	Baked otolith, dorsal arm, 30x magnification under reflected light	
7	Baked otolith, dorsal arm, 40x magnification under reflected light	

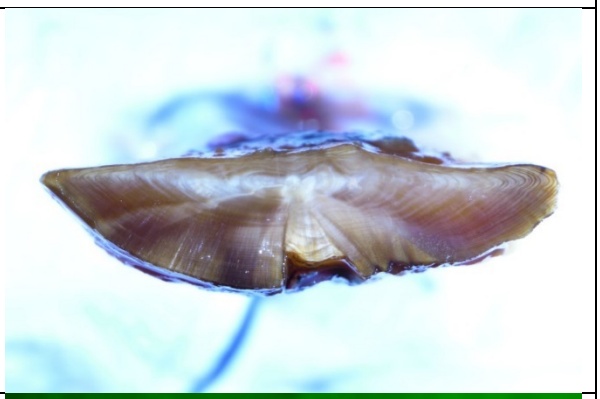
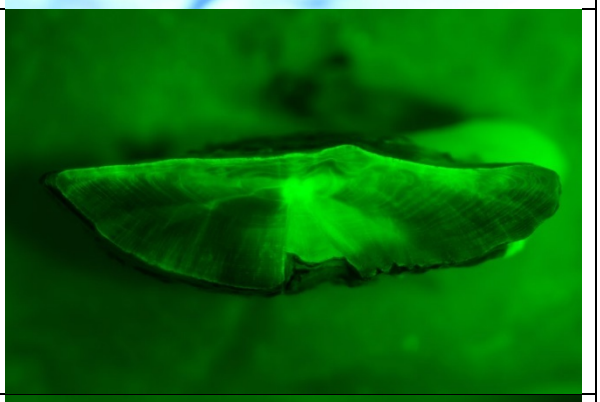
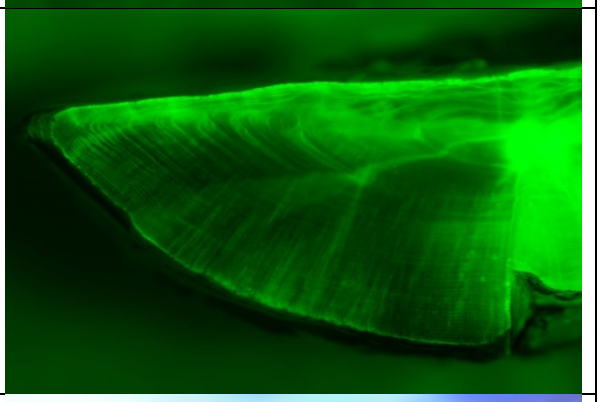
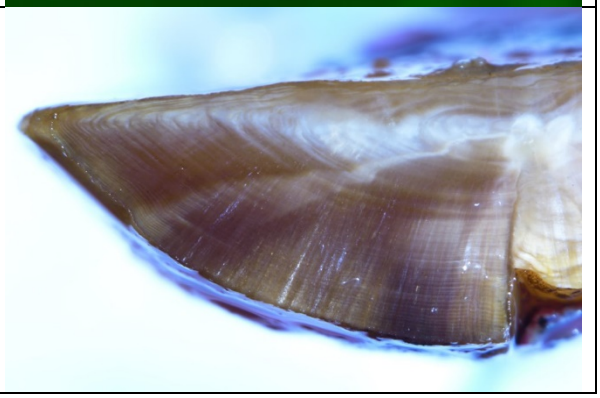
8	Baked otolith, ventral arm, 40x magnification under reflected light	
9	Baked otolith, ventral arm, 50x magnification under reflected light	
10	Baked otolith, dorsal arm, 50x magnification under reflected light	

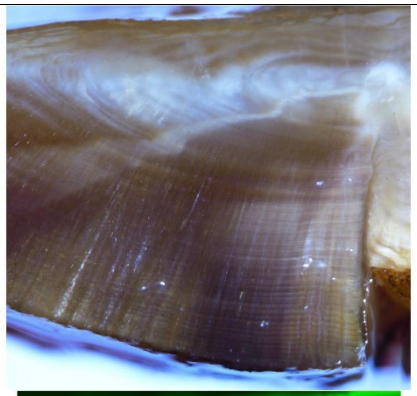
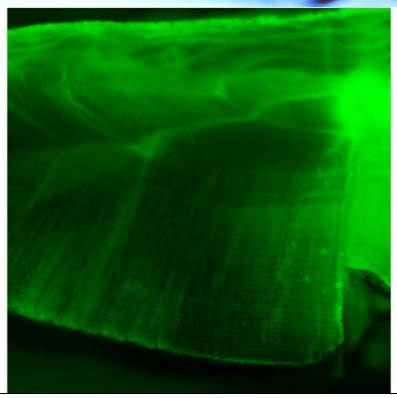
Prior to imaging, all whole undamaged otoliths were weighed to the nearest 0.1 mg for potential inclusion of otolith weight data as data inputs to the CNN models.

2.2.2 Snapper

Break-and-burn prepared otoliths were mounted in plasticine and coated with immersion oil to enhance the series of alternating light and dark zones discernible in the burnt section, following Walsh et al. (2014). Six image types were collected for snapper (Table 2).

Table 2: Examples of images taken for use in age estimation models for snapper (continued on next page).

Image number	Description	Example
1	Break-and-burn prepared otolith, whole face, 10x magnification under reflected light	
2	Break-and-burn prepared otolith, whole face, 10x magnification under ultraviolet light	
3	Break-and-burn prepared otolith, ventral arm, 20x magnification under ultraviolet light	
4	Break-and-burn prepared otolith, ventral arm, 20x magnification under reflected light	

5	Break-and-burn prepared otolith, ventral arm focused on area adjacent to sulcal groove, 40x magnification under reflected light.	
6	Break-and-burn prepared otolith, ventral arm focused on area adjacent to sulcal groove, 40x magnification under ultraviolet light.	

For image types 5 and 6, images of the largest otoliths were captured across two overlapping areas: one encompassing material near the otolith core and a second encompassing material extending to the otolith edge. These two images were then stitched together using Microsoft's Image Composite Editor tool, and padding (i.e., extra pixels) was added to the background to ensure that images were of a consistent size and otoliths were at a consistent relative scale.

2.3 Image modification

A variety of image modification steps, outlined below, were conducted prior to inputting the images to the age estimation algorithm.

2.3.1 Background subtraction

Background subtraction involved removing the background from the focal otolith image (Figure 3) (see also Appendix 1). Background subtraction was conducted using the Clipping Magic software (www.clippingmagic.com; Cedar lakes Ventures, Inc.). Background subtraction was trialled on hoki image types 4, 5, 6, and 8, and snapper image types 4 and 5. Resulting clipped portable network graphic (.png) files were converted back to jpgs for input to the ageing algorithms.

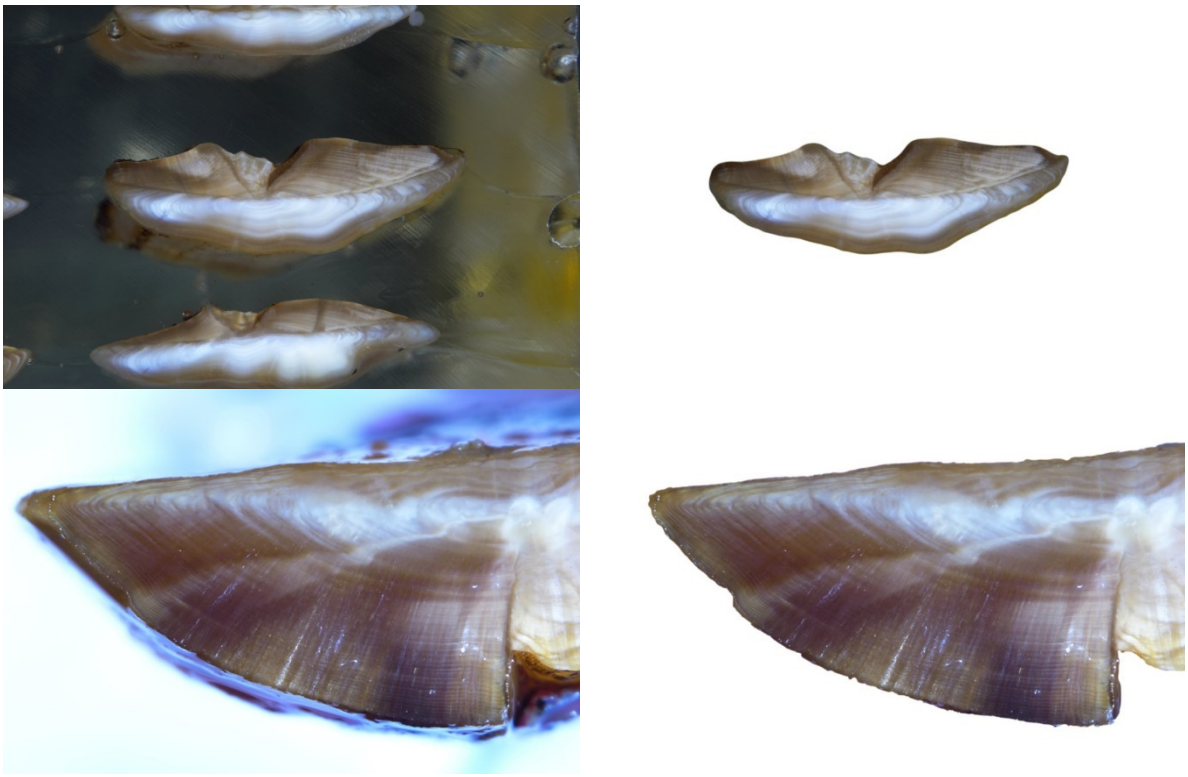


Figure 3: Example of background subtraction for a hoki type 4 image (top) and snapper type 4 image (bottom).

2.3.2 Background subtraction via masking

The background subtraction process detail in Section 2.3.1 resulted in a small loss of information, with images resized from their original dimensions of 4908 x 3264 pixels to 3552 x 2361 pixels. To assess any potential impact of this reduction, a further test was conducted. Here, the clipped otolith image was used as a mask to extract the background from the original otolith image. This enabled the background to be removed from the original otolith image whilst retaining the image's original dimensions (i.e., 4908 x 3264 pixels). Models run using this output were compared against the 3552 x 2361 pixel background-subtracted images.

2.3.3 Image binarisation

The clipped otolith products generated in Section 2.3.1 were used to create binary versions of the original otolith images (Figure 4). This was performed using the R package *imager* (Barthelme 2021) and involved overlaying the original image with the clipped png and setting all pixels in the original image where the transparency channel in the png files was < 0.01 to 0, and all pixels where the transparency channel was ≥ 0.01 to 1.

Figure 4: Example of hoki otoliths (image type 4) having undergone binarisation. The original image is provided on the left for comparison.

2.3.4 Image cropping

Before being fed to the neural network for analysis, each image is rescaled to 299 x 299 pixels (see Section 2.4). To reduce the loss of information caused by this process, each image was cropped to its minimum bounding box (Figure 5), using custom built code in R v 3.6.1. Otolith images were then padded to ensure they matched the dimensions of the largest bounding box (2973 x 938 pixels) to retain the relative scale of each otolith.

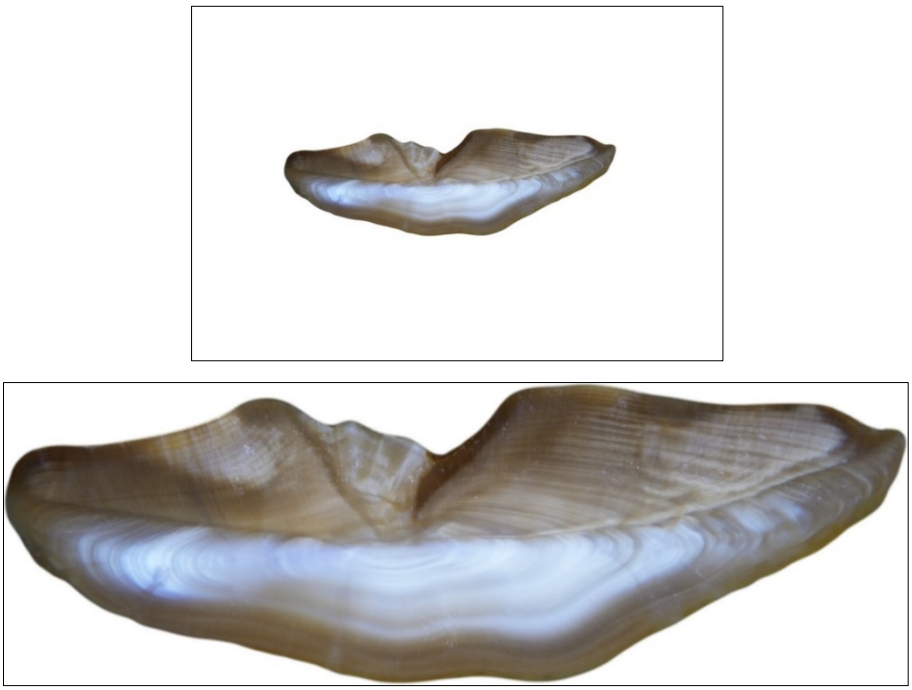


Figure 5: Example of a hoki type 4 image cropped to its minimum bounding box. The original image is shown above for comparison.

In addition, a range of transformations were conducted on each image within the age estimation model framework (see Section 2.4.4).

2.4 Deep learning algorithms

2.4.1 Overview

As given by Moore et al. (2019), ageing algorithms were developed in Python using the TensorFlow Machine Learning library (Abadi et al. 2015) and its Keras application programming interface (API). Since the code for Moore et al. (2019) was written in 2018 and the field of computer vision using deep learning approaches is evolving fast, the latest stable versions available at the beginning of this project were used, which were Python v 3.7 and TensorFlow v 2.3.0. The deep learning algorithm used in the current study followed the same transfer learning process as described by Moore et al. (2019) (Figure 6):

1. loading the Inception V3 CNN (Szegedy et al. 2016) pre-trained on the ImageNet dataset;
2. replacing the classifier layers with a new set of layers adapted for the task of estimating otolith age (i.e., a regressor instead of the original classifier);
3. replacing the Cross-Entropy loss function used in classification algorithms with a Mean Squared Error (MSE) loss function, as commonly used for regression; and
4. retraining the revised machine learning model on the image datasets.

Mean squared error was calculated as the average of the squared differences between the predicted and actual values. The lower the MSE, the better the model is performing. The squaring means that larger differences between model age predictions and human estimates result in more error than smaller differences.

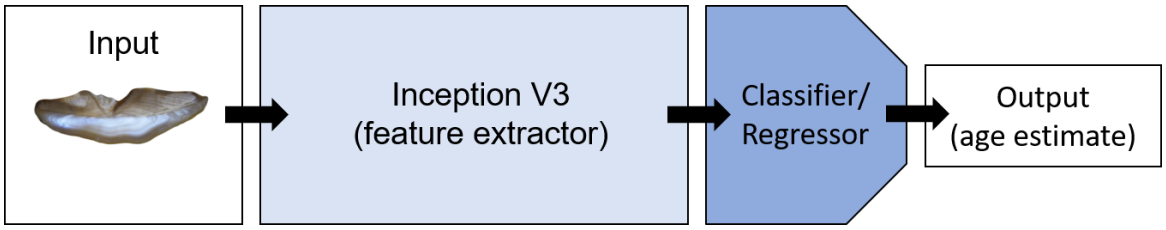


Figure 6: Flow diagram of the neural network architecture used in this study (adapted from Moore et al. 2019).

2.4.2 Datasets

The otolith images were used to train and evaluate the models, with the agreed age for each otolith used as each model's target variable. Developing an efficient machine learning model and accurately evaluating its performance on independent otolith image datasets (generalisation performance) requires careful selection and split of the original dataset. In the first instance, independent models for each type of image were developed, that is, a dataset for training a model consisting of a single image type. Tests were also conducted using models trained on multiple image types (=composite models).

For each model, the image input datasets were split into training, validation, and test subsets (Figure 7). Approximately ~10% of the full dataset was split and set aside for the evaluation of the final model (= test subset). The age distribution of both the hoki and snapper datasets decayed with age. For example, for hoki there were approximately equal numbers for hoki for ages 2–14, with fewer individuals in the youngest and older age classes. Accordingly, this split was stratified by age, so that the final evaluation could be done on all ages available. The remaining ~90% of the image dataset was split (stratified by age) into a training subset and a validation subset, leading to a training/validation/test split of 80/10/10%. The training subset was used to minimise the loss function, and the validation subset was to evaluate the performance of the model (at any stage of the development) on unseen data. Because model development includes tuning hyper-parameters to maximise performance (on the validation subset), the use of separate subsets for model evaluation during development (i.e., the validation subset) and after completion of the development (i.e., the test subset) ensures that the final evaluation is unbiased and more indicative of generalisation performance than the performance on the validation subset.

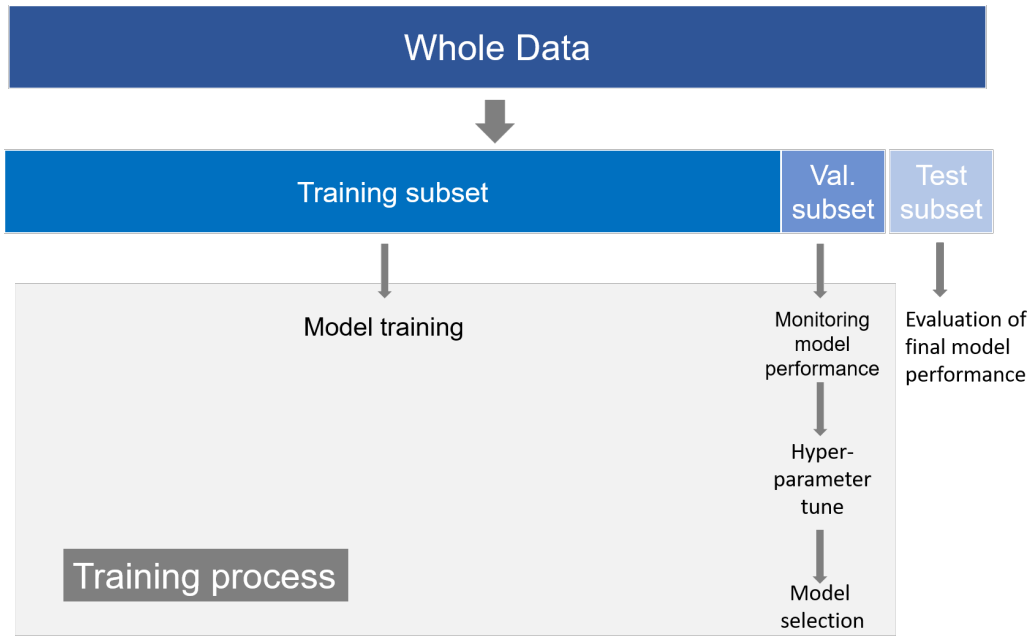


Figure 7: Outline of the split into training, validation (val.) and test data subsets undertaken in the current study.

2.4.3 Model architecture

The architecture and parameters of the final models were optimised to maximise the performance of the model on the validation dataset (*hyper-parameter tuning*).

After loading the Inception V3 CNN, the last layers were removed and replaced with the following sequence of layers:

- a global average pooling 2D layer;
- a dense (fully connected) layer of 1024 units and rectified linear activation functions;
- a dropout layer, with a dropout rate of 0.2; and
- a regressor layer instantiated as a fully connected layer with a single unit and no (i.e., linear) activation function.

In a deep CNN such as Inception V3, the earlier layers tend to capture the gross patterns in the input image while increasingly fine details are captured with increasingly later layers in the network. Accordingly, the common practice in transfer learning is to *freeze* the earlier layers, i.e., prevent the retraining of their parameters, to maintain their existing gross pattern-recognising capabilities. In this study, following the results obtained by Moore et al. (2019), the input layers and the first several blocks of the Inception V3 model were frozen (that is, up to layer 249 out of 315). The neural network had a total of 23 901 985 parameters, of which 13 214 081 were trainable.

2.4.4 Training

Model training for both hoki and snapper was achieved with the Adaptive Movement Estimation (Adam) optimiser (Kingma & Ba 2015), or the Nesterov-accelerated Adaptive Movement Estimation (Nadam) optimiser (Dozat 2016). Adam is a Stochastic Gradient Descent (SGD) method that is based on adaptive estimation of first-order and second-order moments, seeking to minimise the mean square error (*loss function*) between predicted and 'true' age (i.e., that from the human reader). Nadam is an extension of the Adam version of gradient descent that incorporates Nesterov momentum and can often result in improved performance. The Adam parameters, after hyper-parameter tuning, were a learning rate of 1.0×10^{-3}.

Complex models such as a deep CNNs tend to overfit the training data (i.e., poor generalisation performance despite a good fit to the training dataset) when datasets are small, such as in this study. To prevent this outcome, the common method of image augmentation was used, whereby random transformations are applied to the input images to simulate a different training dataset at each epoch. The transformations included:

- "rotation_range": 45;
- "width_shift_range": 0.1;
- "height_shift_range": 0.1;
- "horizontal_flip": True;
- "vertical_flip": True;
- "fill_mode": "reflect".

The Inception V3 CNN uses images of size 299 x 299 as input. All otolith images were rescaled and reduced from their original size of 4908 x 3264 (for the raw images) or the image size specified above as part of the image augmentation process.

Model training was carried out in batches of size 32, over a total of 3000 epochs. At the end of each epoch, the model was evaluated over the validation dataset (without image augmentation, nor dropout) to estimate the model's generalisation performance. An early stopping callback was used to stop the training if this performance had not improved after 750 consecutive epochs.

Since the training of neural networks is highly computing-intensive and parallelisable, the multiprocessing capabilities of the Keras *fit* method that implements the training (8 workers) was used. The training algorithms were run using the CUDA/CUDnn v 10.1 libraries to perform the computations on an NVIDA P100 Graphical Processing Unit off a node of the NIWA/NeSI High Performance Computing Facility.

2.4.5 Model performance evaluation

At the end of an instance of model training, a single measure of performance is obtained: the MSE (loss) on the validation dataset. Additional performance metrics were produced at the end of model training to assist with model evaluation. First, the MSE over the training dataset (without augmentation, nor dropout) was calculated to estimate the final model's performance on the training data. The gap between loss on the validation dataset and training dataset allows diagnosis of cases of overfitting. In addition, the final predicted ages were rounded, and the percentage of correctly predicted ages (percent agreement, PCA), as well as the percentage of predicted age being correct within ±1 year were evaluated, for both the training and validation dataset.

2.5 Ageing performance

Predicted ages from test subsets of candidate models for hoki (model exp_20210524-194903_HOK_55) and snapper (model exp_20210501-054836_SNA_44) were used to evaluate the potential effects on ageing performance. Differences between the human age estimates and model predictions were assessed using the coefficient of variation (CV, Chang 1982), and the index of average percent error (IAPE, Beamish & Fournier 1981). Greater precision is achieved when CV and IAPE are minimised (Beamish & Fournier 1981, Campana et al. 1995). Frequency plots of differences in age estimates between human readers and model predictions, as well as age bias plots modified from Campana et al. (1995), were constructed to detect any differences in age and the presence of any systematic bias in age estimates.

A single age frequency was generated for each of the human age estimates and model predicted ages for the test subsets. Potential differences in age frequency distributions between readers (human vs. CNN model) were tested using Kolmogorov-Smirnov (K-S) tests.

Reproducibility of growth parameter estimates was assessed by constructing separate von Bertalanffy growth function (VGBF) curves from the human age estimates and model predictions for the test subsets. The form of the VBGF used to model length-at-age data was:

$$L_t = L_\infty[1 - e^{-k(t-t_0)}]$$

where L_t is the length-at-age t, L_∞ is the hypothetical asymptotic length, k is the growth coefficient, and t_0 is the hypothetical age at which fish would have zero length. Resulting growth function curves and parameters were compared using likelihood ratio tests (LRT, Kimura 1980).

3. RESULTS

3.1 Hoki

3.1.1 Models run on unmodified images

Models run on 'raw' (i.e., unmodified) hoki images produced largely similar results across image types. For the test data subsets, MSEs ranged from 3.91 (for image type 1) to 2.33 (for image type 8), and correct agreement with human readers ranged from 25.2% (image types 6 and 7) to 35.5% (image types 2 and 3) (Table 3).

3.1.2 Models run on modified images

Models run using 'clipped' (i.e., background removed) images showed variable results relative to those using raw images. For hoki image types 4 and 5, removing the background improved the MSE for the test subsets. In contrast, for those image types taken at higher magnification (i.e., image types 6 and 8), MSE values for the test subsets were slightly higher for model runs using images with the background removed than when using the raw images (Table 4).

The model run on binarised type 4 images achieved MSE values for the test subset that were slightly lower (better) than those of the model run on raw images (4.08 vs. 4.20), but higher than that achieved by the model run on clipped images (2.71). Overall, percent agreement, and the percent of individuals within ± 1 of the human estimates, from the model run on binary versions of hoki image type 4 were largely comparable, if not slightly improved, than those from the models using raw and clipped images (Table 4).

3.1.3 Composite models

A single composite model was run for hoki, incorporating one image type of a whole otolith (image type 3) and one image type of a bake-and-embed prepared otolith (image type 5). This model achieved MSE values that were slightly higher than model runs using a single image type for the validation dataset, and MSE values that were generally comparable to other models for the test dataset. The percent agreement, and the percent of individuals within ± 1 of the human estimates, from this model were generally lower than those resulting from models run on single image types, at 13.1% and 41.1%, respectively (Table 5).

Table 3: Results for models run on unmodified hoki images. MSE = mean squared error, PCA = percent agreement, val. = validation subset.

Model run name	Image type	Best epoch	MSE train	MSE validation	MSE test	PCA val.	PCA ± 1 val.	PCA test	PCA ±1 test
exp_20210503-150642_HOK_1	1	254	0.7416	2.3769	3.9083	32.71	76.64	28.97	66.36
exp_20210504-100656_HOK_2	2	143	1.1285	2.6992	3.0252	28.97	68.22	35.51	77.57
exp_20210507-111019_HOK_3	3	54	1.5675	2.2938	2.8905	25.23	66.36	35.51	69.16
exp_20210528-213855_HOK_4	4	652	0.2476	2.5687	4.1975	29.91	71.96	20.56	59.81
exp_20210523-090901_HOK_5	5	1 136	0.1336	2.6583	2.9288	29.91	70.09	28.04	71.03
exp_20210507-221237_HOK_6	6	983	0.1075	2.1469	3.0789	31.78	70.09	25.23	65.42
exp_20210509-063913_HOK_7	7	497	0.3209	2.2065	3.4827	30.84	75.70	25.23	70.09
exp_20210510-063137_HOK_8	8	844	0.2117	2.7126	2.3325	31.78	66.36	34.58	68.22
exp_20210511-124921_HOK_9	9	286	0.4264	3.3253	2.8029	28.04	61.68	28.97	70.09
exp_20210512-074621_HOK_10	10	558	0.2298	2.3864	2.7124	36.45	76.64	33.64	71.03

Table 4: Results for models run on modified hoki images. MSE = mean squared error, PCA = percent agreement, val. = validation subset. Note results for models run on the unmodified images using the same model settings are shown for comparison.

Model run name	Image type	Modification	Best epoch	MSE train	MSE validation	MSE test	PCA val.	PCA ± 1 val.	PCA test	PCA ± 1 test
Run using optimiser Adam:										
exp_20210528-213855_HOK_4	4	Unmodified	652	0.2476	2.5687	4.1975	29.91	71.96	20.56	59.81
exp_20210529-231430_HOK_444	4	Background removed	741	0.4816	2.5755	2.7142	21.50	65.42	19.63	64.49
exp_20210530-044224_HOK_44444	4	Binarised	501	2.9949	2.2589	4.0757	40.19	71.03	32.71	64.49
exp_20210523-090901_HOK_5	5	Unmodified	1 136	0.1336	2.6583	2.9288	29.91	70.09	28.04	71.03
exp_20210524-194903_HOK_55	5	Background removed	399	0.341	2.9298	2.7933	28.97	62.62	29.91	72.90
exp_20210507-221237_HOK_6	6	Unmodified	983	0.1075	2.1469	3.0789	31.78	70.09	25.23	65.42
exp_20210522-042312_HOK_66	6	Background removed	1 361	0.1703	2.3786	3.1522	32.71	63.55	31.78	70.09
exp_20210510-063137_HOK_8	8	Unmodified	844	0.2117	2.7126	2.3325	31.78	66.36	34.58	68.22
exp_20210522-220202_HOK_88	8	Background removed	480	0.5292	3.4065	3.2751	31.78	68.22	30.84	69.16
Run with optimiser Nadam, removed kernel regularization in Dense layer:										
exp_20210425-194729_HOK_4	4	Unmodified	88	0.7432	2.8451	4.314	24.30	62.62	25.00	56.48
exp_20210426-063313_HOK_44	4	Background removed	373	0.8924	2.9538	2.8825	34.58	64.49	38.32	71.96
exp_20210426-130601_HOK_444	4	Masked	845	1.1678	2.7365	2.9223	28.04	67.29	34.58	72.90
exp_20210427-065721_HOK_5	5	Unmodified	212	0.4475	1.8184	3.7822	33.64	76.64	33.33	73.15
exp_20210427-201306_HOK_55	5	Background removed	867	0.1551	2.6733	2.6913	35.51	73.83	38.32	76.64

Table 5: Results for models run incorporating multiple image types per individual. MSE = mean squared error, PCA = percent agreement, val. = validation subset.

Model run name	Image types	Modification	Best epoch	MSE train	MSE validation	MSE test	PCA val.	PCA ± 1 val.	PCA test	PCA ± 1 test
Run using optimiser Adam:										
exp_20210618-171743_HOK_353_535	3 and 5	Unmodified	134	2.974	4.3765	3.1153	16.98	41.51	13.08	41.12

3.2 Snapper

3.2.1 Models run on unmodified images

Models run on break-and-burn prepared snapper otolith images performed relatively poorly, with MSE values on the test subsets ranging from ~9.1 to 35.2 for those run using the optimiser Adam, and ~9.5–10.5 for models run using the optimiser Nadam (Table 6). Models run using the optimiser Adam resulted in ages of 13.2–26.4% of samples being correctly predicted, whereas models run using the optimiser Nadam resulted in a correctness in age prediction of 28.3% and 37.7% for images types 4 and 5, respectively (Table 6).

3.2.2 Models run on modified images

Where trialled, removing the background had a minor effect on age estimates for snapper, with a slight improvement in MSE, percent agreement, and percent of reads within one year for the validation subset, but yielded negligible improvement across these metrics for the test subset (Table 7).

3.2.3 Composite models

A single composite model was run for snapper, incorporating image types 4 and 5. This model achieved MSE values of 47.26 for the validation dataset, and 21.97 for the test dataset (Table 8).

Table 6: Results for models run on unmodified snapper images. MSE = mean squared error, PCA = percent agreement, val. = validation subset.

Model run name	Image type	Best epoch	MSE train	MSE validation	MSE test	PCA val.	PCA ± 1 val.	PCA test	PCA ± 1 test
Run using optimiser Adam									
exp_20210604-141910_SNA_1	1	23	8.0534	14.0635	22.2042	9.62	42.31	13.21	43.40
exp_20210604-214808_SNA_2	2	385	2.8064	13.9299	13.2863	21.15	48.08	26.42	43.40
exp_20210605-104400_SNA_3	3	261	1.7705	17.0455	9.1052	23.08	51.92	16.98	47.17
exp_20210605-214852_SNA_4	4	1 436	1.9634	12.7627	12.3892	19.23	40.38	26.42	49.06
exp_20210606-195051_SNA_5	5	509	1.5638	14.2187	15.9396	17.31	46.15	19.23	46.15
exp_20210607-091557_SNA_6	6	1 032	1.0545	25.0676	35.2176	21.15	51.92	19.23	40.38
Run using optimiser Nadam									
exp_20210423-134008_SNA_4	4	1 237	0.5707	11.7820	9.4561	26.92	50.00	28.30	49.06
exp_20210424-134125_SNA_500	5	357	1.9928	14.2146	10.5343	19.23	42.31	37.74	62.26

Table 7: Results for models run on modified snapper images. MSE = mean squared error, PCA = percent agreement, val. = validation subset.

Model run name	Image type	Modification	Best epoch	MSE train	MSE validation	MSE test	PCA val.	PCA ± 1 val.	PCA test	PCA ± 1 test
Run using optimiser Adam										
exp_20210605-214852_SNA_4	4	Unmodified	1 436	1.9634	12.7627	12.3892	19.23	40.38	26.42	49.06
exp_20210501-054836_SNA_44	4	Background removed	397	2.32	10.3769	10.4901	25.00	67.31	28.30	49.06
Run using optimiser Nadam										
exp_20210423-134008_SNA_4	4	Unmodified	1 237	0.5707	11.7820	9.4561	26.92	50.00	28.30	49.06
exp_20210424-074839_SNA_44	4	Background removed	701	1.0298	11.1959	9.2238	28.85	69.23	20.75	49.06

Table 8: Results for models run incorporating multiple image types per individual. MSE = mean squared error, PCA = percent agreement, val. = validation subset.

Model run name	Image types	Modification	Best epoch	MSE train	MSE validation	MSE test	PCA val.	PCA ± 1 val.	PCA test	PCA ± 1 test
exp_20210425-044016_SNA_44_500	4 (clipped) and 5	Background removed (4) and unmodified (5)	216	54.1861	47.2613	21.9712	19.23	28.85	9.43	20.75

3.3 Ageing performance

3.3.1 Hoki

Age estimates from model exp_20210524-194903_HOK_55 were used to explore the effect of using the age estimated from the CNN model estimates for the generation of biological parameters in place of those from human readers. The mean CV of the CNN predicted ages was 7.07% (Figure 8), and the IAPE was 5.00%. This was reduced to a CV of 6.72% when a single outlier was removed (Figure 8), which also reduced the IAPE to 4.75%. Overall, there was a slight tendency for the model to underestimate the ages of fish in the oldest age classes (Figure 8, Figure 9). It should be emphasised, however, that there were fewer old fish in both the test and training datasets (Figure 1).

No significant difference was evident between the age frequency distributions for human estimates and predicted ages from the model (K-S test, D= 0.0374, P=1.0; Figure 9). Similarly, no significant difference was evident in any of the VBGF parameters between growth curves generated from human estimates and model predictions (Figure 10, Table 9, Table 10).

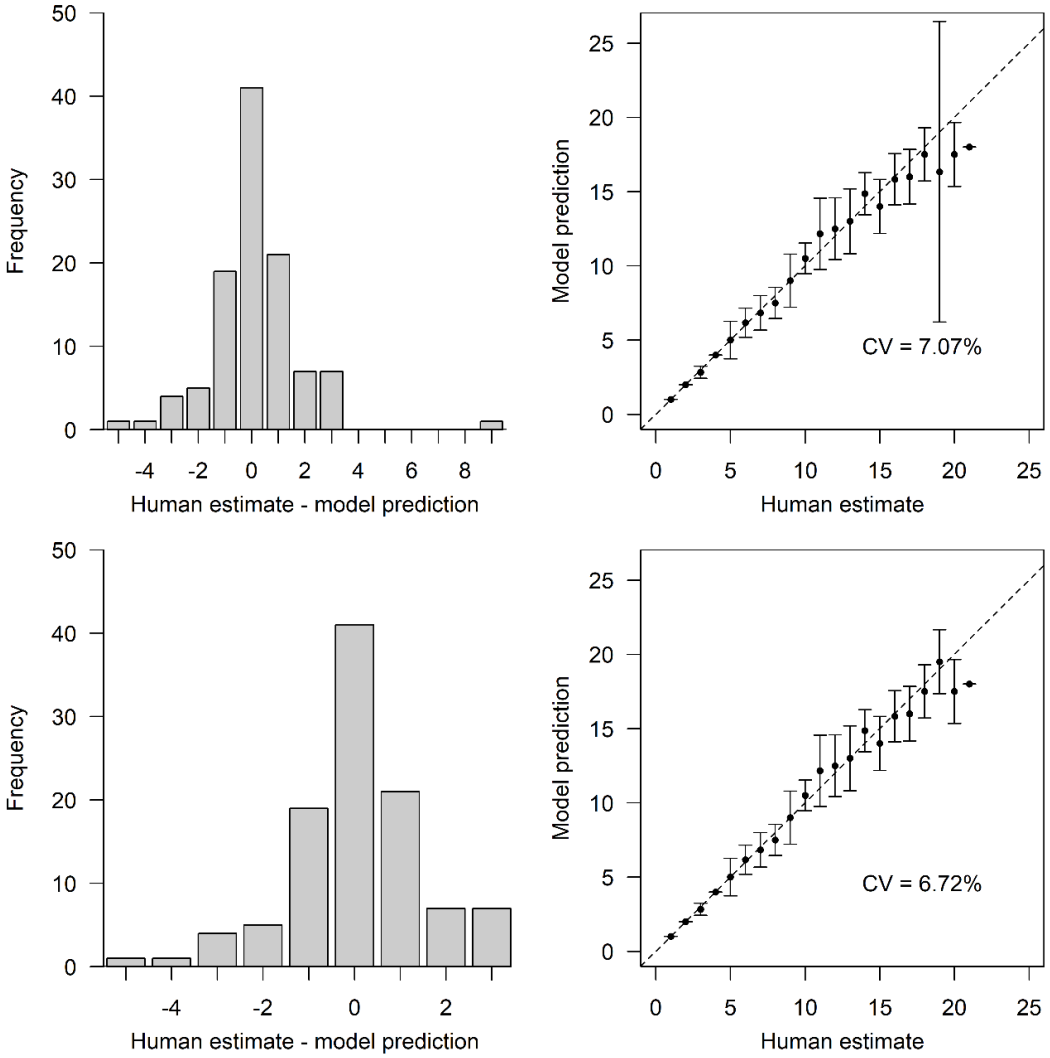


Figure 8: Distribution of age frequency differences and age bias plots from the test subset of the candidate CNN model for hoki (model exp_20210524-194903_HOK_55). Results are presented for all individuals in the test dataset (n=107; top row) and excluding one outlier (bottom row).

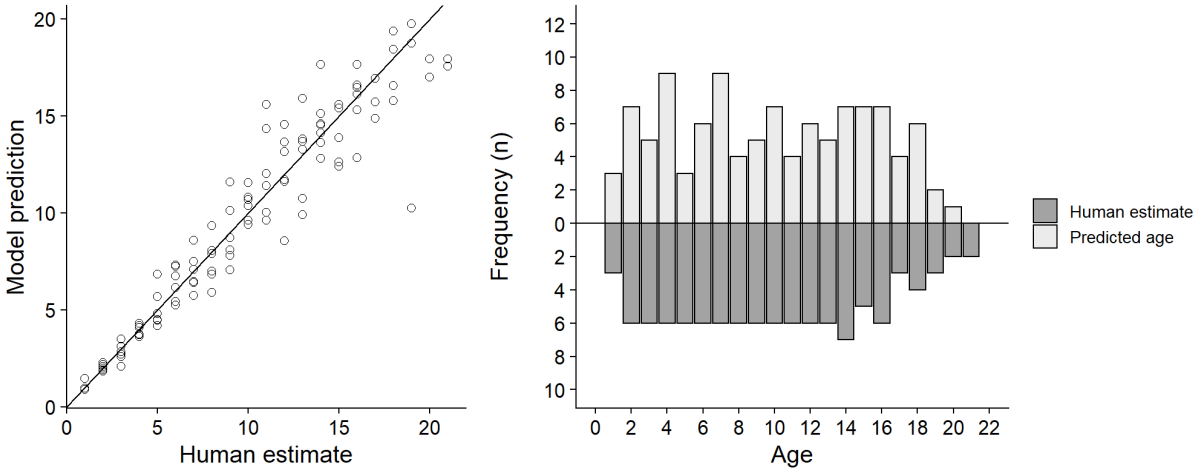


Figure 9: **Scatterplot of predicted ages vs. human estimates (left) and age frequency distributions for hoki from the human estimates and predicted ages (right) from the test subset of the candidate CNN model.**

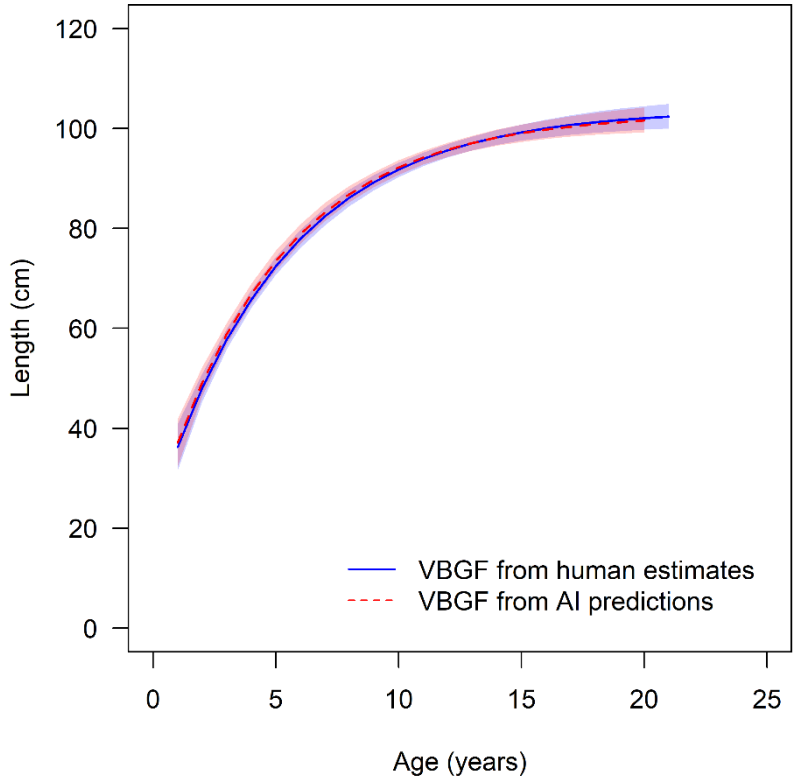


Figure 10: **von Bertalanffy growth curves for hoki generated from age predictions from the test subset of the candidate CNN model and corresponding human estimates.**

Table 9: **von Bertalanffy growth function parameters (± standard error, SE) for hoki from models based on human age estimates and CNN predicted ages.**

Parameter	Human age estimates	Model predicted ages
L_∞	103.86 (±1.71)	103.06 (±1.76)
k	0.19 (0.02)	0.20 (0.02)
$t0$	-1.24 (0.32)	-1.23 (0.34)

Table 10: Results from likelihood ratio tests comparing von Bertalanffy growth parameter estimates generated from human age estimates and the test subset of the candidate CNN model for hoki.

Test	χ^2	P value
L_{inf} (Human) vs. L_{inf} (Model)	0.57	0.450
k (Human) vs. k (Model)	0.19	0.663
t_0 (Human) vs. t_0 (Model)	0.14	0.708
All	0.85	0.837

3.3.2 Snapper

The mean CV of the CNN predicted ages from the test subset of model exp_20210501-054836_SNA_44 was 10.89% (Figure 11), and the IAPE was 7.70%. Overall, there was a slight tendency for the model to overestimate the age relative to human estimates (Figure 11). Again, it should be emphasised that there were very few old fish in the test and dataset, and in the dataset that the model was trained on (Figure 2).

No significant difference was evident between the age frequency distributions for human estimates and predicted ages from the model (K-S test, D=0.0943, P=0.97; Figure 12). Similarly, no significant difference was evident in any of the VBGF parameters between growth curves generated from human estimates and model predictions (Figure 13, Table 11, Table 12).

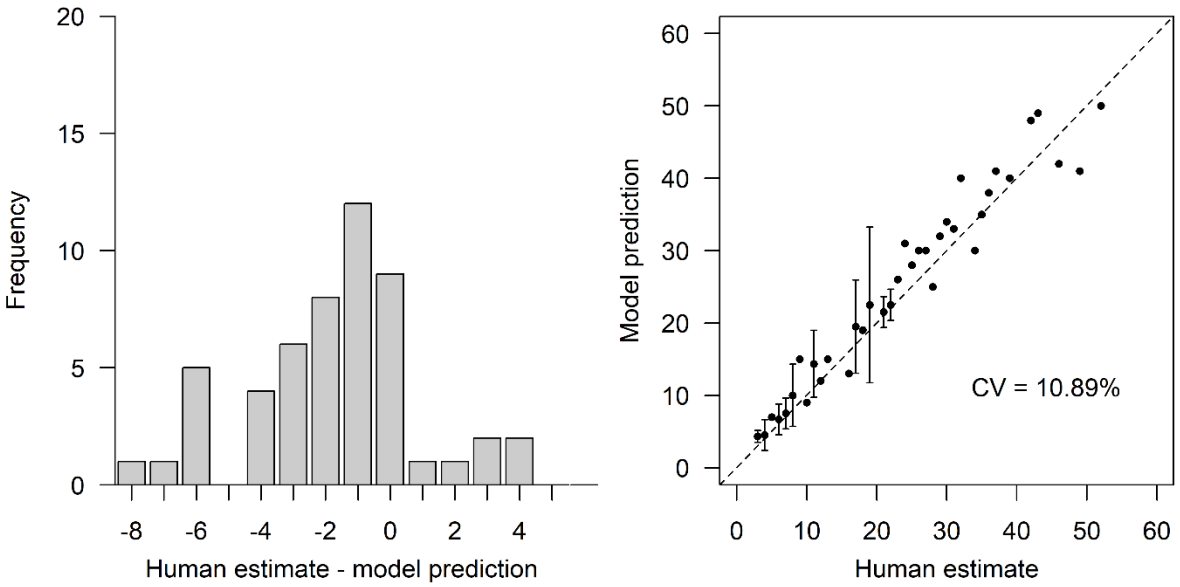


Figure 11: Distribution of age frequency differences and age bias plots from the test subset of the candidate CNN model for snapper (model exp_20210501-054836_SNA_44).

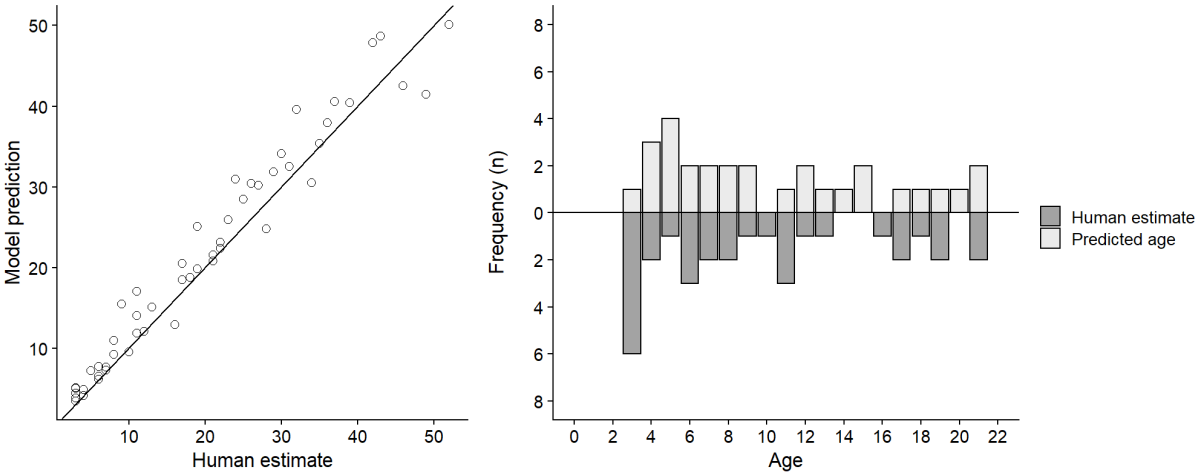


Figure 12: Scatterplot of predicted ages vs. human estimates (left) and age frequency distributions for snapper from the human estimates and predicted ages (right) from the test subset of the candidate CNN model.

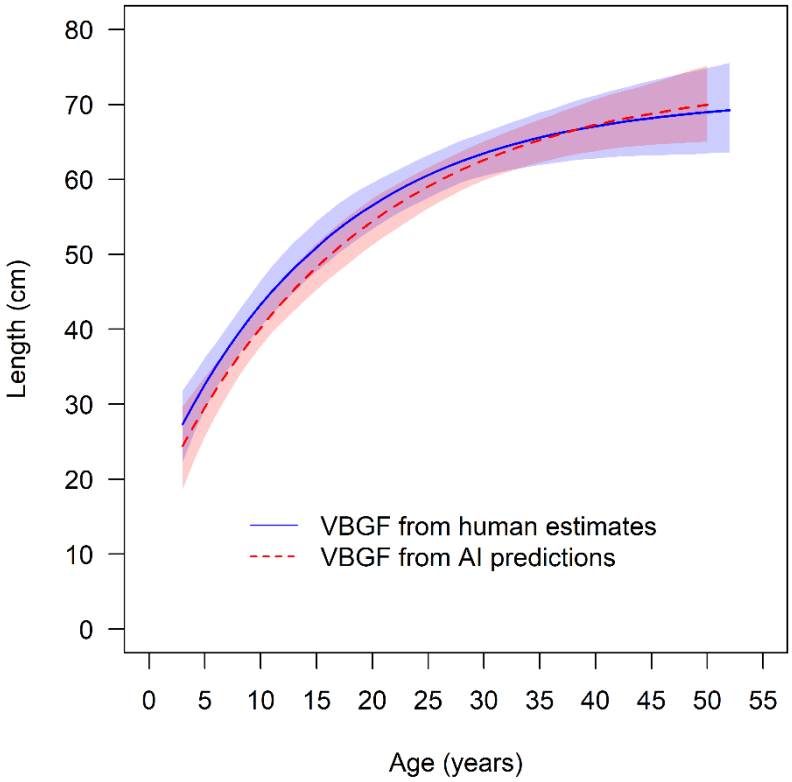


Figure 13: von Bertalanffy growth curves for snapper generated from age predictions from the test subset of the candidate CNN-model and corresponding human estimates.

Table 11: von Bertalanffy growth function parameters (± standard error, SE) for snapper from models based on human age estimates and CNN predicted ages.

Parameter	Human age estimates	Model predicted ages
L_∞	71.05 (4.79)	73.62 (5.50)
k	0.06 (0.02)	0.06 (0.02)
$t0$	-4.46 (2.24)	-4.28 (2.30)

Table 12: Results from likelihood ratio tests comparing von Bertalanffy growth parameter estimates generated from human age estimates and the test subset of the candidate CNN model for snapper.

Test	χ^2	P value
L_{inf} (Human) vs. L_{inf} (Model)	0.01	0.920
k (Human) vs. k (Model)	0.04	0.841
t_0 (Human) vs. t_0 (Model)	0.23	0.632
All	0.54	0.910

4. DISCUSSION

The processing and reading of fish otoliths for age estimation is a time-consuming and costly procedure (Worthington et al. 1995, Francis & Campana 2004), which can impede the implementation of routine ageing programmes required for monitoring and assessment. Results of this study build upon those of previous research (e.g., Moen et al. 2018, Moore et al. 2019, Ordoñez et al. 2020, Politikos et al. 2021) that demonstrate the significant potential for using deep learning CNN-based approaches to predict ages of fish from otolith images.

The precision of age estimations achieved here for hoki (CV=7.07% for the test subset of the examined hoki model) compare favourably with those cases in the primary literature where similar methodologies have been applied. For example, Moen et al. (2018) achieved a mean CV of 8.89% for their CNN-based model applied to images of whole otoliths of Greenland halibut (*Reinhardtius hippoglossoides*). Interestingly, this was despite Moen et al. (2018) utilising a much larger training dataset (8165 images) than used in the current study.

For hoki, the CNN model slightly underestimated the age of the oldest fish (i.e., those greater than 18 years) relative to human readers. Underestimation of oldest age classes is commonly observed in CNN-based studies of age estimation (e.g., Moen et al. 2018) and likely results from reduced numbers of old fish relative to younger age classes in both training and testing datasets, as well as finer-scale deposition of opaque and translucent material with increasing age, coupled with loss of information through rescaling of image dimensions for input to the CNN (Moen et al. 2018, Moore et al. 2019, Ordoñez et al. 2020). Nevertheless, comparisons of key biological parameters generated from model predictions showed no significant difference when compared with those generated from age estimation by human readers, demonstrating little overall effect of this bias, or other minor age prediction differences, on parameter estimation.

Age estimates derived from the CNN-based models for snapper in the current study were less accurate than the results for hoki, and for snapper prepared as thin sections by Moore et al. (2019), for which an MSE of 1.2 and overall PCA of 46.7% were achieved for the test subset. This likely reflects the break-and-burn method of preparation of the samples examined and associated challenges with imaging. The otolith samples used commonly had small chips and other damage to the cut surface, with non-uniform ground axes, and differed in the degree of burning and subsequent colouration, resulting in inconsistent size and shape, and problematic interpretability between samples. Although a human reader can ignore this information, the trial using binarised images for hoki and other published works (e.g., Ordoñez et al. 2020, Politikos et al. 2021) suggest that the CNN model likely uses size and shape attributes of the otolith, in addition to patterns in translucent and opaque material, to derive age estimates. While selecting 'optimal' otoliths (i.e., those in perfect, or near-perfect condition) may have improved the result, this study opted instead to ensure that the technique was developed on samples that are representative of those available for ageing. It is recommended that future development and implementation of routine ageing of snapper via machine learning be conducted on thin section-prepared otoliths, consistent with the validated approach (Francis et al. 1992), and as used for routine ageing in other jurisdictions (e.g., South Australia by Fowler et al. 2016).

4.1 Next steps / recommendations for future research

These results, and particularly those for hoki, highlight the considerable potential for using CNN models to derive age estimates from otolith images. Below, the key next steps for improving and implementing both the image capture and age estimation components of this work are presented.

1. Evaluate the hoki model(s) against the current otolith reference collection. When introducing new readers into an ageing programme, best practice is to ensure they have been tested against a reference set of otoliths. The primary role of a reference set is to monitor ageing consistency (and accuracy) over both the short and long term, particularly for testing long-term drift, as well as consistency among age readers (Campana 2001). The current reference collection for hoki in New Zealand consists of 480 individual otoliths covering a range of fish lengths and collection seasons, with approximately equal coverage between stocks and sexes (Horn & Sutton 2017). Horn & Sutton (2017) consider that IAPE values of less than 5% when tested against the reference set are reasonable for hoki. Accordingly, it is strongly recommended that age predictions from candidate ageing algorithms for hoki be tested against the reference collection to evaluate whether they meet this benchmark.

2. Improve / automate image capture. The capture of images was a costly and time-consuming component of the current project and is a considerable impediment to implementing the approach for routine age estimation. Approaches to optimise image capture should be investigated as a priority. For those species for which routine ageing is performed via thin otolith sections on microscope slides (e.g., trevally *Pseudocaranx dentex*, tarakihi *Nemadactylus macropterus*), one such approach may be to trial digital slide scanning technologies for automating image capture. Slide scanning machines can capture high-resolution images with high throughput (i.e., multiple slides at a time). Such a trial should first investigate the appropriateness of estimating age from captured images (e.g., by comparing ages derived on-screen from images against prior age estimates from experienced human readers) before using the images in a CNN-framework.

3. Further develop the image segmentation model. Results of the present study revealed that models based on images with backgrounds removed in most cases outperformed those with backgrounds retained. Of the background removal approaches trialled (see Appendix 1), the image segmentation approach holds considerable promise for further development, because it was found to be generally robust and flexible enough to detect fine-scale differences between background and foreground components. Further development of this model is required, including generation of both larger training and test datasets, running the segmented products from the CNN age estimation model, and trialling the performance of models trained on different image types.

4. Resolve issues around resizing of the images within the CNN model. The default size of the images used in the Inception V3 is 299 x 299 pixels. Increasing (or retaining the original size) the size of the input images may provide greater resolution of structural patterns, particularly of the outermost bands in older specimens. However it would require considerable model training to replace the feature extractor layers from the Inception V3 CNN (Szegedy et al. 2016), and more computing resources to develop each model. Determination of the size of images will be a balance between providing good quality images to the CNN and the limitations of computational resources.

5. Improve understanding of the features the models are trained on. For stakeholders to accept the age estimations from a CNN model, and for the resulting age estimates to be used for fisheries management, confidence needs to be built through some level of decision understanding. It is especially important to verify that the model does not learn biased prediction rules based on some artefacts related to the acquisition of the images or other non-age-related features, such as background material. Although significant investment was made in removing noise and

arbitrary information from the images (e.g., background removal), due to resource constraints this study was unable to explore in detail the model fitting characteristics of the CNN algorithms, and determine which features the CNNs were using to determine age. Further research to verify which features the developed models are trained on is necessary. Understanding decisions of deep learning algorithms is commonly achieved by visualisation approaches. Layer-wise relevance propagation (Bach et al. 2015) is one potential approach to visualise the decisions of the deep learning algorithms that has been previously applied to otolith images for age estimation purposes (e.g., Ordoñez et al. 2020). This approach aims to assign the importance of an input pixel to the overall output prediction score by back-propagating a relevance score encoding the information about the model's decision (Ordoñez et al. 2020).

6. Further develop an integrated approach using multiple image types and other data inputs (e.g. otolith weight). Multi-task learning (MTL), a nascent subfield of machine learning, provides one potential approach for integrating different data types. Here, instead of focusing on a single task, multiple auxiliary tasks are used simultaneously, allowing for sharing of information between networks (Politikos et al. 2021). Accordingly, an auxiliary task of the prediction of fish age from otolith weight, other variables such as 2-dimensional otolith area, or potentially other image types (i.e., in a composite, multi-image model approach), could be introduced to the CNN. A recent study by Politikos et al. (2021) used a MTL approach to better estimate age of red mullet (*Mullus barbatus*) by introducing as an auxiliary task the prediction of fish length from otolith images, improved overall age prediction, and proved effective at identifying older age classes.

5. ACKNOWLEDGMENTS

We thank the staff and contractors involved with the collection and processing of otoliths used in this study. Debbie Hulston and Keren Spong kindly assisted with the extraction of otoliths from sample archives. Kameron Christopher and Maxime Rio provided recommendations for algorithms and analyses and reviews of Python model code. Julian Maclaren (Nelson AI Institute) provided useful discussion on image modification and augmentation. Richard Saunders, Jeremy McKenzie, and Richard O'Driscoll provided constructive comments on an earlier draft of this report. This work was funded by Fisheries New Zealand under project SAM2019-02.

6. REFERENCES

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jozefowicz, R.; Jia, Y.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Schuster, M.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.; Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One 2015*: e0130140.

Barthelme, S. (2021). imager: Image Processing Library Based on 'CImg; R package version 0.42.8. https://CRAN.R-project.org/package=imager

Beamish, R.J.; Fournier, D.A. (1981). A method for comparing the precision of a set of age determinations. *Canadian Journal of Fisheries and Aquatic Sciences 38*: 982–983.

Bose, A.P.H.; Adragna, J.B.; Balshine. S. (2017). Otolith morphology varies between populations, sexes and male alternative reproductive tactics in a vocal toadfish *Porichthys notatus*. *Journal of Fish Biology 90:* 311–325.

Campana, S.E. (2001). Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. *Journal of Fish Biology 59*: 197–242.

Campana, S.E.; Annand, M.C.; McMillan, J.I. (1995). Graphical and statistical methods for determining the consistency of age determinations. *Transactions of the American Fisheries Society 124*: 131–138.

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*: 30–38.

Chang, W.Y.B. (1982). A satistical method for evaluating the reproducibility of age determination. *Canadian Journal of Fisheries and Aquatic Sciences 39*: 1208–1210.

Dozat, T. (2016). Incorporating Nesterov Momentum into Adam. International Conference for Learning Representations Workshop Paper, 2016. Available at https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ

Fablet, R.; Le Josse, N. (2005). Automatic fish age estimation from otolith images using statistical learning. *Fisheries Research 72*: 279–290.

Fowler, A.J.; McGarvey, R.; Carroll, J.; Feenstra, J.E.; Jackson, W.B.; Lloyd, M.T. (2016). Snapper (*Chrysophrys auratus*) fishery. Fishery Assessment Report to PIRSA Fisheries and Aquaculture. South Australian Research and Development Institute (Aquatic Sciences), Adelaide. *SARDI Research Report Series No. 930*. 82 p.

Francis, R.I.C.C.; Campana, S.E. (2004). Inferring age from otolith measurements: a review and a new approach. *Canadian Journal of Fisheries and Aquatic Sciences 61*: 1269–1284.

Francis, R.I.C.C.; Paul, L.J.; Mulligan, K.P. (1992). Ageing of adult snapper (*Pagrus auratus*) from otolith annual ring counts: Validation by tagging and oxytetracycline injection. *Australian Journal of Marine and Freshwater Research 43*: 1069–1089.

Horn, P.L.; Sullivan, K.J. (1996). Validated aging methodology using otoliths, and growth parameters for hoki (*Macruronus novaezelandiae*) in New Zealand waters. *New Zealand Journal of Marine and Freshwater Research 30*: 161–174.

Horn, P.L.; Sutton, C.P. (2017). Age determination protocol for hoki (*Macruronus novaezelandiae*). *New Zealand Fisheries Assessment Report 2017/13.* 22 p.

Kimura, D.K. (1980). Likelihood methods for the von Bertalanffy growth curve. *Fishery Bulletin 77*: 765–776.

Kingma, D.P.; Ba, J. (2015). Adam: a method for stochastic optimization. 3rd International Conference for Learning Representations, San Diego, 2015. Available at https://arxiv.org/abs/1412.6980

Mackay, K.A.; George, K. (2017). Age. Database documentation for the Ministry of Primary Industries ageing database. (Unpublished NIWA Fisheries Data Management Database Documentation Series held by the National Institute of Water and Atmospheric Research library, Wellington.) 59 p.

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*: 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 2019/58*. 23 p.

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*: e0235013.

Parmentier, E.; Boistel, R.; Bahri, M.A.; Plenevaux, A.; Schwarzhans, W. (2018). Sexual dimorphism in the sonic system and otolith morphology of *Neobythites gilli*. *Journal of Zoology 30*: 274–280.

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.

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*: 73–82.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. (2016). *In*: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 2818–2826. 27–30 June 2016, Las Vegas, NV, USA.

Walsh, C.; Horn, P.; McKenzie, J.; Ó Maolagáin, C.; Buckthought, D.; Sutton, C.; Armiger, H. (2014). Age determination protocol for snapper (*Pagrus auratus*). *New Zealand Fisheries Assessment Report 2014/51*. 33 p.

Welch, T.J.; van den Avyle, M.J.; Betsill, R.K.; Driebe, E.M. (1993). Precision and relative accuracy of striped bass age estimates from otoliths, scales, and anal fin rays and spines. *North American Journal of Fisheries Management 13*: 616–620.

Worthington, D.G.; Fowler, A.J.; Doherty, P.J. (1995). Determining the most efficient method of age determination for estimating the age structure of a fish population. *Canadian Journal of Fisheries and Aquatic Sciences 52*: 2320–2326.

Zhu, X.; Wastle, R.J.; Howland, K.L.; Leonard, D.J.; Mann, S.; Carmichael, T.J.; Tallman, R.F. (2015). A comparison of three anatomical structures for estimating age in a slow-growing subarctic population of lake whitefish. *North American Journal of Fisheries Management 35*: 262–270.

APPENDIX 1: Trial for removal of image backgrounds

Preliminary trials revealed that age estimates were significantly improved if backgrounds were removed from input images. Accordingly, a series of background removal methods were trialled to determine which approach provided the optimal result. These trials were conducted on hoki image types 1 (i.e., whole otolith imaged at 4x magnification) and 4 (full face of bake-and-embed prepared otolith imaged at 11x magnification) (Figure A1.1).

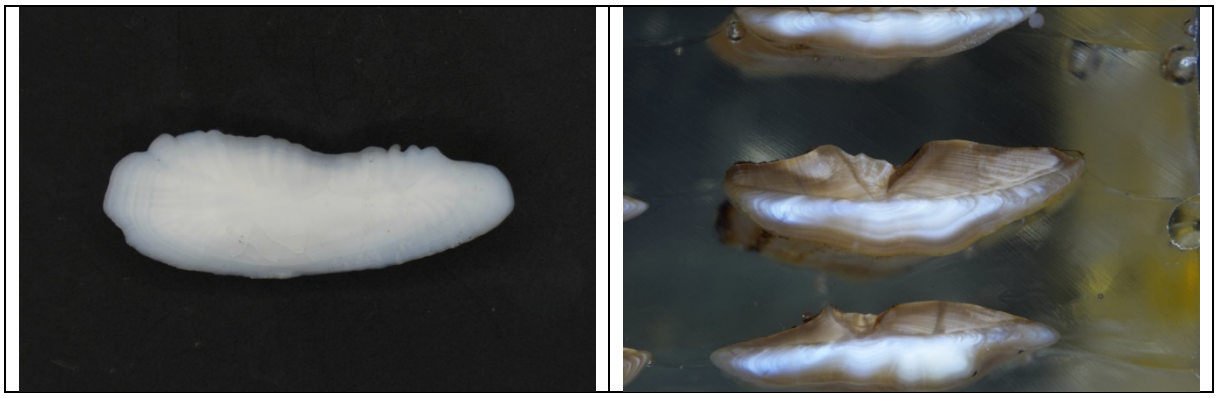


Figure A1.1: Base images used in background removal trials. Left: hoki image type 1; right: hoki image type 4.

Background subtraction via ImageJ

The subtract background function of ImageJ was trialled as a first step in isolating the focal otolith from the image background. This approach performed well for the whole otolith image, but failed to differentiate the focal otolith from neighbouring images, non-focal material (e.g., material beyond the cut surface) and overall background in the bake-and-embedded image (image type 4; Figure A1.2).

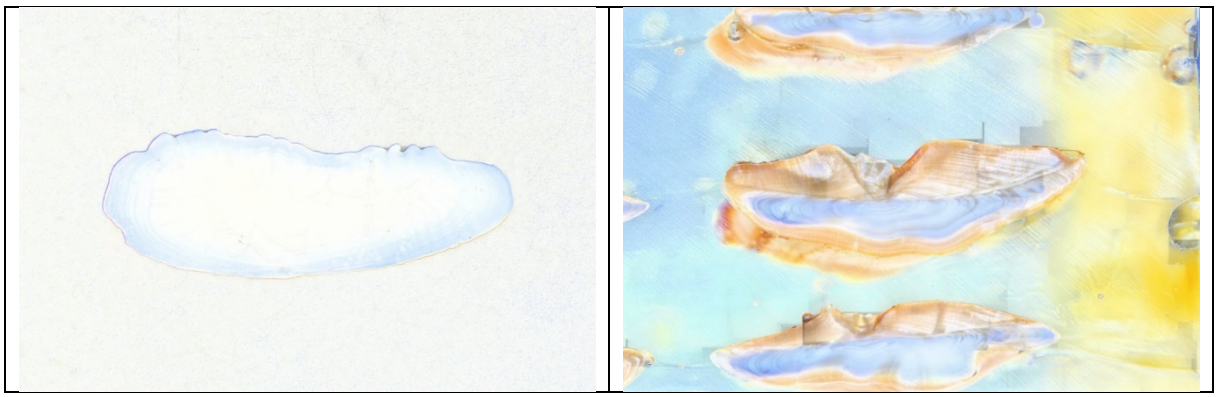


Figure A1.2: Images with background subtracted using subtract background function in ImageJ.

Background subtraction via *ShapeR* package

The R package *ShapeR* (Libungan & Pálsson 2015) includes a function that detects the outline of a focal object. It is most commonly used for extracting otolith shape information for investigations of species identification and stock structure (Libungan & Pálsson 2015). If the outline of the focal otolith could be obtained, a shape mask could be developed to mask the background from the focal otolith in original raw images.

The *detect.outline* function performed well for whole otoliths, accurately identifying the outlines in most cases (Figure A.1.3). However, it failed to accurately detect the required outline in bake-and-embed prepared otolith sections (Figure A.1.3), often being unable to differentiate the focal otolith from

background material, or detecting outlines from adjacent otoliths, even across the full range of threshold settings.

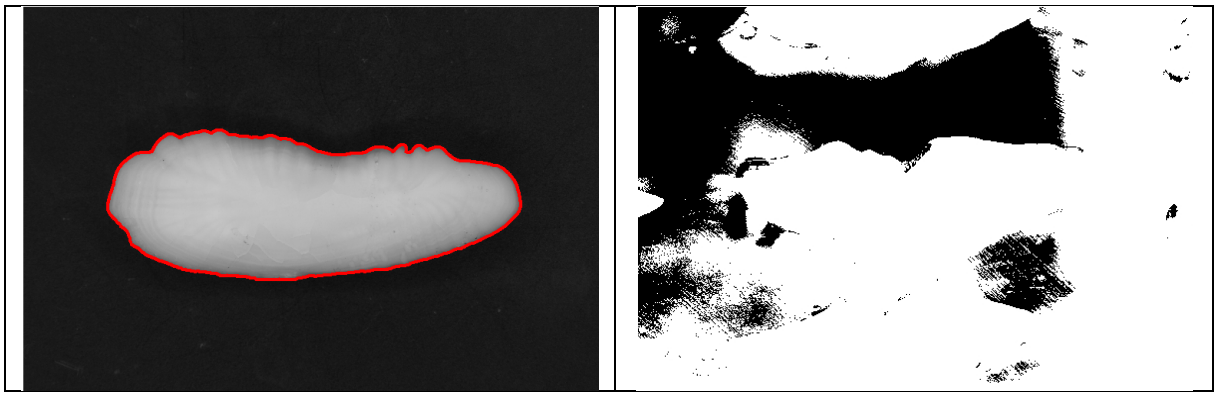


Figure A1.3: Otolith outlines detected using *detect.outline* function from *ShapeR* package in R.

Background subtraction via *Tensorflow & Keras* (Image segmentation)
Here, a u-net algorithm was used to develop a neural network to automatically identify the focal otolith from surrounding material. U-net is a multi-class semantic segmentation algorithm, allowing classification of each pixel to an object (in this case the otolith of interest or the background). The algorithm was developed in R using the packages *tensorflow* (Allaire & Tang 2021), *keras* (Allaire & Chollet 2021), and *platypus* (Maj 2021).

The model was trained on a subset (n=1057) of paired original and binary mask images for hoki image type 4 (Figure A1.4). The binary masks were generated by binarising the resulting product generated via the Clipping Magic tool described in Section 2.3, using the R package *imager* (Barthelme 2021). Due to time constraints, training was based on images reduced to 512 x 512 pixels, and the training model was run for 100 epochs. Resulting predicted masks (Figure A1.5) were overlaid on the original images to extract the focal otolith, remove the image background, and evaluate model performance.

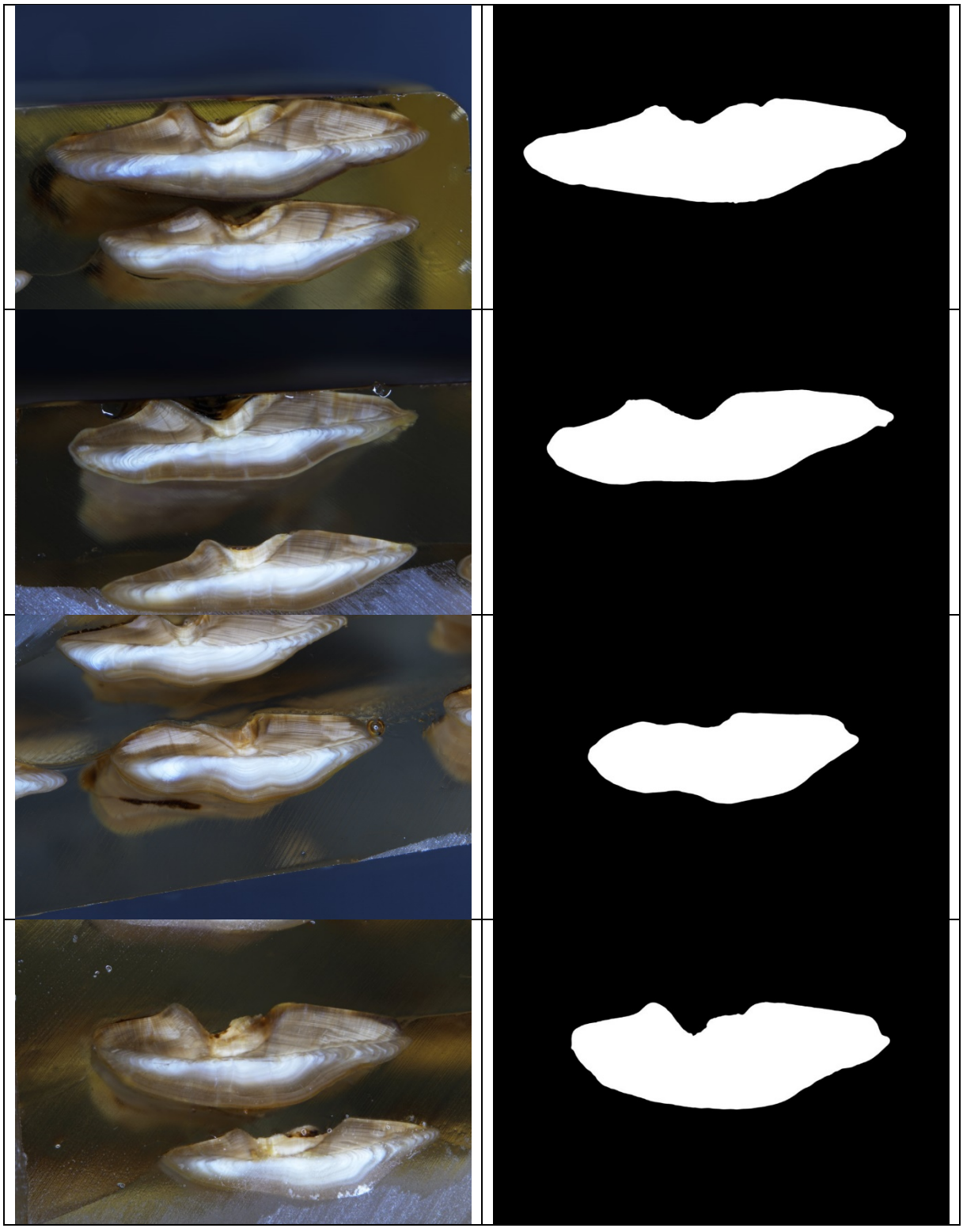


Figure A1.4: Examples of original image and binary mask pairs used to train the image segmentation model.

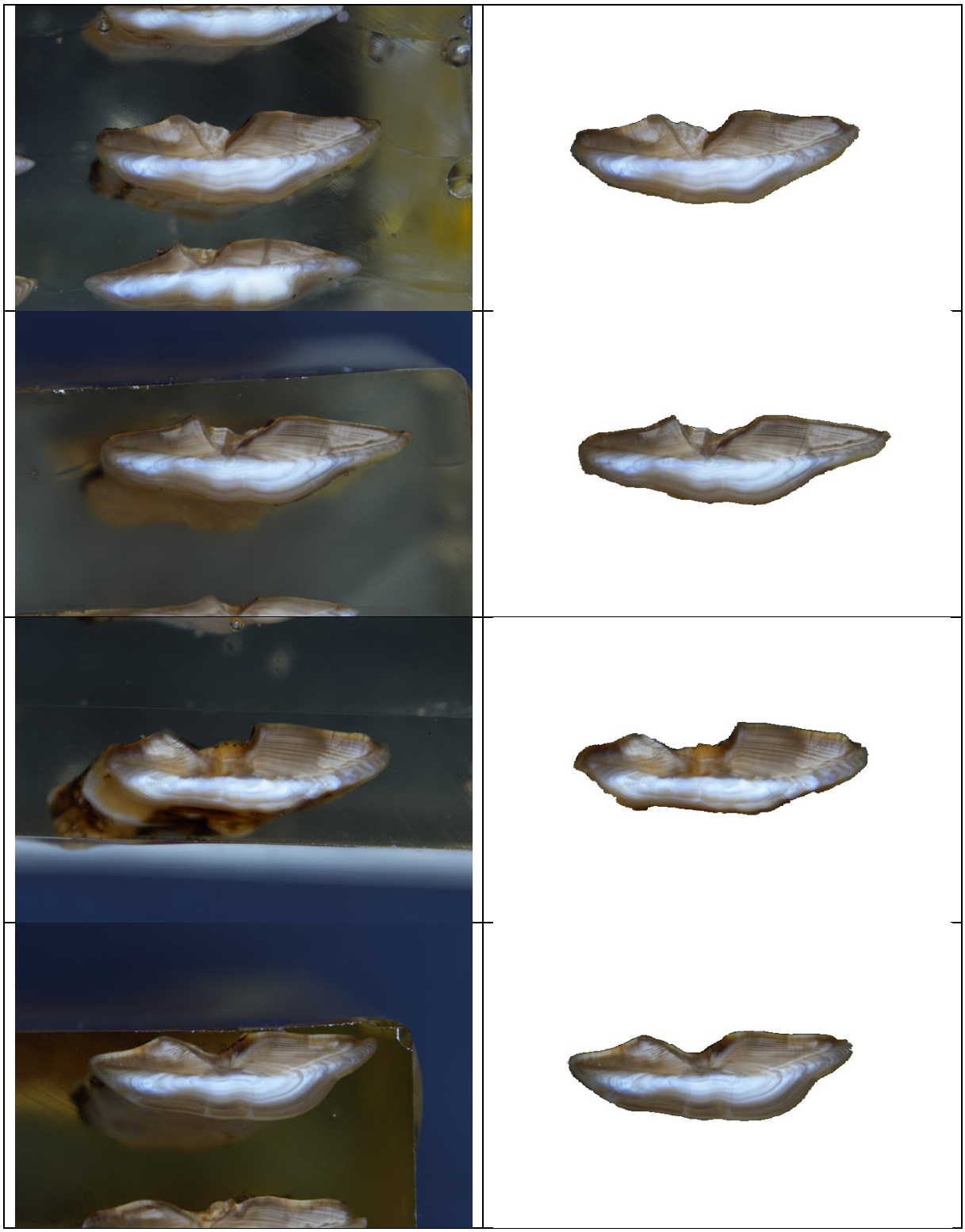


Figure A1.5: Examples of predicted segmented images from the image segmentation model (right column). The original images are shown in the left column for comparison.

References

Allaire, J.J.; Chollet, F. (2021). keras: R Interface to 'Keras'. R package version 2.4.0. https://CRAN.R-project.org/package=keras

Allaire, J.J.; Tang, Y. (2021). tensorflow: R Iinterface to 'Tensorflow'. R package version 2.4.0. https://CRAN.R-project.org/package=tensorflow

Barthelme, S. (2021). imager: Image Processing Library Based on 'CImg; R package version 0.42.8. https://CRAN.R-project.org/package=imager

Libungan, L.; Pálsson, S. (2015). ShapeR: An R package to study otolith shape variation among fish populations. *PLoS ONE 10*: e0121102.

Maj, M. (2021). platypus: Tools for Computer Vision in R. R package 0.1.1.