

Age interpretation of cod otoliths using deep learning

Endre Moen^{1*}, Rune Vabø¹, Szymon Smoliński², Côme Denechaud¹, Nils Olav Handegard¹, Ketil Malde^{1,3},

1 Institute of Marine Research, Bergen, Norway

2 Department of Fisheries Resources, National Marine Fisheries Research Institute,
Kołłątaja 1, 81-332 Gdynia, Poland

3 Department of Informatics, University of Bergen, Norway

* endre.moen@hi.no

Abstract

The age ~~structure composition~~ of fish populations ~~is a central input~~ plays a crucial role in stock management ~~, and a valuable data in most~~ and provides valuable information for biological studies. Fish age ~~can be inferred by~~ is typically estimated by manually counting annual increments in otoliths. ~~Typically, fish age is estimated manually by experts by visual interpretation of otolith structures,~~ but this process is ~~time and resource-consuming~~ prone to age reader bias and requires considerable time and resources. Consistency between readers and labs are also a challenge. In this study, we ~~developed~~ a machine learning framework for ~~automating the~~ age estimation of Atlantic cod (*Gadus morhua*) based on multi-exposure otolith images based on images of broken otoliths (N = 5150). We apply the two architectures (~~Images of broken otoliths from Atlantic cod (*Gadus morhua*) were used as the test case. We used EfficientNetV1 and EfficientNetV2~~) using with three and two different model sizes ~~from each family, comparing the performance of different models and their ensembles~~, respectively, from each type and compared the performance. The average accuracy ~~of the~~ 17 and mean squared error when comparing predictions and the manually read ages for the tested models was 72.7% and ~~the average MSE was~~ 0.284, respectively.

The models' accuracy for one and two years old individuals was over 90% and no systematic bias in the age predictions across age groups was observed. **Models**
~~EffieientNet~~ The best models were ~~EfficientNet~~ B4 and EfficientNet B6 ~~with the lowest image exposure were the top-scoring models using images taken with low exposure times.~~ After an exhaustive search, a maximum accuracy of 78.6% was achieved with an ensemble consisting of six models. The tested models strongly correlate in terms of predictions. Variations in percentage agreement between age classes showed similar patterns in both CNN-based predictions and human readers, with generally decreasing agreement with age. While percentage agreement from CNN-based predictions was often lower than for human experts, it remained within or close to the range of percentage agreement observed across readers. Our results demonstrate that the use of deep learning techniques in the analysis of otoliths has potential for facilitating automation. When developing ~~the framework for automatic age estimation frameworks for age estimation using machine learning~~, we recommend ~~including~~ ~~EfficientNet B4 architectures~~ ~~models~~ as they are quick to train and perform well. Ensemble approaches are also recommended if sufficient computational resources are available, as they can give increased accuracy and lower variance of the predictions. We find that under-exposed images tend to perform better than ~~other exposures. Our study can serve as a baseline for the future development and operationalization of CNN models and the inclusion of automatic otolith-based age interpretations in the biological data collection routines.~~ ~~longer exposures.~~ In contrast to previous studies trained on otolith sections requiring a time-consuming preparation, we used broken otoliths that require no processing before imaging. This shows the potential for more resource-efficient training framework providing near *at-sea* age estimates.

Introduction

1 Introduction

Knowledge of fish age structure is central to ~~the study of fish and stock dynamics~~ ~~fisheries science and stock dynamic modelleing~~. It informs on population growth and mortality and is one of the main criteria used for determining the health of

exploited populations and monitoring the effects of selective fishing (Hidalgo et al., 47
2011; Brunel and Piet, 2013). Changes in the age distribution can track significant 48
changes in population structure, such as the appearance of a particularly strong 49
year-class (Reglero and Mosegaard, 2006), or the gradual truncation of older age classes 50
as selective fishing mortality removes larger individuals (Siskey et al., 2016). Hard 51
structures such as scales and otoliths are used worldwide as one of the primary sources 52
of fish age estimates, due to their ability as natural physiological and environmental 53
recorders to form regular, temporally resolved growth increments **at the annual levels** 54
(Campana, 2001; Francis and Campana, 2011; Albuquerque et al., 2019). While age is 55
inferred from the “simple” counting of annual increments, the interpretation of this 56
zonation pattern is species or even population-specific (Høie et al., 2009) and is based 57
on precise knowledge of the timing of zone formation and of the correct identification of 58
true and false zones (Panfili et al., 2002). This process, therefore, requires specific 59
expertise and is subject to uncertainties **and bias** in both between-reader precision and 60
“true” age accuracy (Francis and Campana, 2011). Therefore, streamlining, scaling, and 61
increasing the quality of age estimations can improve the reliability of evaluations of fish 62
biology and consequently assessment of stock size and structure (Tyler et al., 1989; 63
Beamish and McFarlane, 1995; Ragonese, 2018). 64

Otolith reading is time and resource-consuming. Training of expert readers can take 65
several years depending on the species, and otoliths often undergo a long processing 66
phase before the final age estimates can be produced (Carbonara and Follesa, 2019). 67
This is particularly true for demersal fish species, like Atlantic cod (*Gadus morhua*), 68
that have large opaque otoliths that typically require time-consuming preparation 69
(Denechaud et al., 2020; Smoliński et al., 2020). These routines vary between **species** 70
and populations and institutes and range from a direct reading of broken otoliths under 71
a magnifying glass, to embedding, thin sectioning, and finally imaging of the sections 72
under a microscope. **All otoliths read in Norway and Russia for the Northeast Arctic** 73
cod population are read on broken otoliths using a magnifying glass. 74

There has been a variety of methods proposed to automatically interpret otoliths, 75
which range from one-dimensional data analysis like intensity transects (Mahé, 2009) to 76
the more recent effort toward developing machine learning (ML) frameworks (Moen 77
et al., 2018; Politikos et al., 2021; Sigurdardóttir et al., 2023). **One of the main** 78

advantages of automation is that the results are reproducible and consistent. When
new models emerge, it is also straight forward to rerun the analysis on the material
and evaluate the impact on the use case, e.g. a fisheries assessment model.

1.1 Deep learning and image analysis

During the last decade, deep learning has become one of the dominating fields of
machine learning where various architectures of deep neural networks are able to learn
to efficiently ~~and~~ identify patterns and structures in various types of data (LeCun et al.,
2015). Within the field of computer vision, deep Convolutional Neural Networks (CNN)
have been widespread ever since ~~(Krizhevsky et al., 2012)~~ Krizhevsky et al. (2012) won
the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition
(Russakovsky et al., 2014). ILSVRC remains the most important benchmark for image
classification with 1.4 million images in the ImageNet training set, and state-of-the-art
CNNs are therefore often targeted to this data set. Many of these CNNs are publicly
available including their trained network weights. It is therefore often useful to use
transfer learning with these pre-trained weights as a starting point. This is especially
true for tasks where relatively little training data is available. For many fish species, age
estimation from images of otoliths represents precisely such a task. InceptionV3
(Szegedy et al., 2015) was modified to predict the age of Greenland halibut
(Reinhardtius hippoglossoides) from otolith images (Moen et al., 2018), and a modified
InceptionV3 was applied to classify otolith images of red mullet (*Mullus barbatus*)
(Politikos et al., 2021). While some state-of-the-art CNNs grew in model size, a recent
CNN architecture called EfficientNet (Tan and Le, 2019) demonstrated that increased
performance could be achieved with smaller model sizes (number of parameters) using a
compound scaling method for network depth, width and image size, resulting in a
family of seven different models with different sizes. This network has been successfully
applied with transfer learning to analyse images of salmon scales (Vabø et al., 2021).
Recently a successor to the EfficientNet architecture, EfficientNetV2 (Mingxing
Tan and, 2021), has been made available.

~~In this study, we develop a~~ The objective of this study is to develop a deep
learning framework for automating the age estimation of Atlantic cod based on

multi-exposure images of broken otoliths. We apply the two EfficientNet family architectures taken with constant illumination and three different exposures. We will test EfficientNetV1 and EfficientNetV2 using three and two different a range of model sizes from each family respectively. We , and we aim to compare the performance of the different models and discuss the use of , including an ensemble of models to improve estimation accuracy. Our study identifies specific features of investigated architectures and suggests the . We aim to provide best practices and strategies to develop for developing CNN frameworks for the predictions of the age of fish based on otoliths. It We further anticipate that this can serve as a baseline for the future development and operationalization of CNN models and the inclusion of automatic otolith-based ML-based otolith age interpretations in the biological data collection routines.

Method and materials

2 Method and materials

2.1 Data Collection

We used a data set sampled from 5150 cod otoliths collected on surveys conducted by the Institute of Marine Research (IMR) in the period 2012-2018 and aged by otolith experts. On each of the surveys, the otoliths were sampled using a random-stratified sampling based on fish length for each trawl station.

Each otolith was broken in the transverse plane and placed on a mount before it was captured by six images with three light exposures and one rotation of 180° (Figure 1). The images were taken with a resolution of 3744×5616 pixels. The image light exposure punctually varied depending on light conditions coming from outside. Light exposure was stored in the metadata of the JPG file. Details can be found in (Myers et al., 2019) and in the data set available at <https://doi.org/10.21335/NMDC-1826273218>.

2.2 Convolutional neural network architecture

Each CNN was trained using transfer learning by loading ImageNet weights. The training images were resized from 3744×5616 pixels to between 380×380 and 528×528

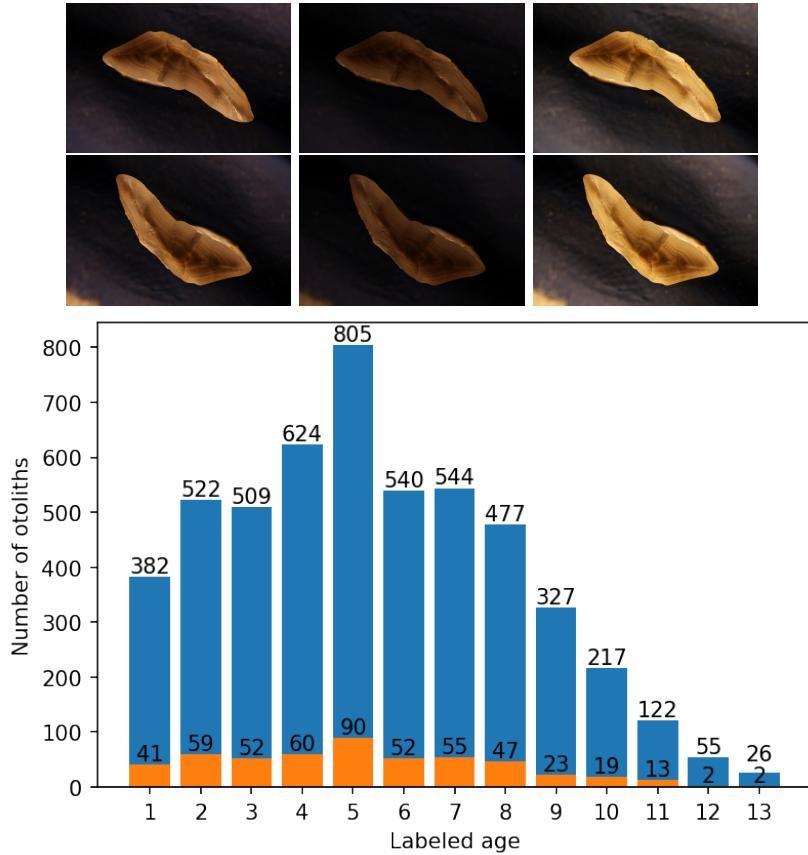


Figure 1. Images of an otolith collected in 2016 from a 6 years old cod (top), taken with medium-, minimum- and max-exposures (upper row), then rotated 180° (lower row). The age distribution of the 5150 otoliths in the training (blue), and the 515 otoliths in the test (orange) set (bottom).

pixels depending on the architecture. The pixel values have a range between 0 and 255, 136 which was normalized to between 0 and 1. Test set predictions were done on images 137 resized to 380×380 and 384×384 pixels. To investigate the effect of exposure and 138 orientation as presented in the image-taking protocol described in (Myers et al., 2019), 139 we also trained on 9-channel images by stacking ~~3 RGB images representing the three~~ 140 color layers from each of the three images representing 141 different lighting exposures. 142 Using Timm (Wightman, 2019), the ImageNet weights were duplicated on the input 143 layer to accommodate all 9 channels. The three images used were of dark, medium, and 144 light exposure.

CNNs were selected based on performance on the ImageNet benchmark and the 145 availability of open-source implementations with pre-trained weights. The CNN models 146 are aimed at classification, while we treated aging as a regression problem (Moen et al., 147

2018; Vabø et al., 2021). The last layer of the CNNs was therefore modified to a linear
148 output. For the EfficientNetV2 family we did this by applying three multi-layer
149 perceptron layers going from 1280 output of the last hidden layer to a dense 256-layer,
150 then a leakyRelu (Xu et al., 2015) layer, then a dense 32-layer, then a leakyRelu layer,
151 and finally a linear output layer. For EfficientNet we only changed the last layer from
152 softmax to a linear output (Figure 8 in supplementary materials).
153

To each fold, we normalized the age on the training set by subtracting the mean and
154 scaling to unit variance. The normalization was then applied to the validation and test
155 sets. Test set predictions were obtained by applying the inverse transform.
156

2.3 Implementation and training

157

EfficientNetV1 B4, B5, and B6 were imported and modified with TensorFlow (Abadi
158 et al., 2016) and Keras (Chollet and others, 2018) software packages in Python.
159

Computation was done using CUDA 11.1 and CuDNN with Nvidia (Nvidia Corp.,
160 Santa Clara, California) A6000 accelerator card with 48 GB of GPU memory and P100
161 cards with 12 GB of GPU memory, EfficientNetV2 Medium, and Large were imported
162 and modified with the PyTorch (Paszke et al., 2019) and Timm (Wightman, 2019)
163 software packages. Computation was done on P100 and RTX 3090 with 24 GB of GPU
164 memory. Pretrained weights for EfficientNet were available from Keras, and pre-trained
165 weights for EfficientNetV2 were available from Timm. The models will be referred to as
166 B4-Min, B4-Middle, B4-Max, B5-Min, Medium-min and so on by combining model
167 name with image exposure.
168

Augmentation ~~was applied to the training set is a commonly used technique to artificially inflate the training data set by applying transforms that modify the input while preserving class.~~ The images were augmented using rotation between 0 and 360 degrees, and reflection by the vertical axis.
169
170
171
172

The cost function used was mean squared error (MSE) while the metric used for
173 evaluating the models and comparing them to expert readers was accuracy. Accuracy
174 was obtained by rounding the ~~floating point number real valued~~ predictions to the
175 nearest integer and ~~comparing the age classification against the true measuring the fraction of otoliths where the age classification matches the labels.~~
176
177

The data set of 5150 otoliths was divided into a training set constituting 90% of the 178 otolith images (4635 otoliths) and a test set of 10% (515 otoliths). To get the most out 179 of a small data set we applied 10-fold cross-validation on the training set. This meant 180 that 10% of the training set was used for validation and 90% (81% of the whole data 181 set) was used for the actual training for each fold. Consequently, The data set is divided 182 into ten parts, and in each iteration (or "fold"), a different part is retained for 183 validation, while the model is trained on the remaining nine parts. In other words, 184 10 different models were trained with a different set of 463 images used for validation in 185 each fold, i.e. each data point participates in the validation set once and in the training 186 set 9 times. Among the 10-fold models, the one with the best MSE was chosen. The 187 best model parameters on the validation set were then used to predict the age on the 188 test set, and the metric for accuracy and MSE were recorded. The test set is chosen at 189 random, while the 10-fold split of the training set is chosen using a stratified k-fold split, 190 which preserves a similar distribution of the whole cross-validation set in each validation 191 set. That means the 463 images in the validation set will have similar age distribution 192 to that of the 4635 images in the cross-validation set. 193

2.4 Hyperparameters

 194

The CNN ~~hyper-parameters~~ hyperparameters (i.e., model parameters that are set in 195 advance, in contrast to model parameters that are learned during training) 196 configurations varied a little between the two families of networks, but were kept the 197 same within the families. Some ~~hyper-parameters~~ hyperparameters that were tuned are 198 batch size, learning rate, k-fold size, weight decay, step size, number of epochs, early 199 stopping, and patience. Some parameters are constrained by the GPU memory, like 200 batch size which was set to 16 for models trained on the A6000 card, and to 8 for the 201 models trained on P100s. 202

EfficientNet used learning-rate with a weight decay scheduler, while EfficientNetV2 203 used Cosine Annealing scheduler (Loshchilov and Hutter, 2016). The training- and 204 validation image size used was not changed, except for EfficientNetV2 Large which uses 205 a smaller validation image size. The exact configuration of each network is available 206 with each network result on the GitHub page of the project 207

(<https://github.com/emoen/Deep-learning-for-regression-of-cod-otoliths>). The
208
~~hyper-parameters~~ [hyperparameters](#) are available in Table 3, and 4 in supplementary
209
information.
210

2.5 [Ensemble learning with averaging](#) 211

Ensemble learning is an algorithm that combines the predictions from multiple models
212
to reach a final prediction and obtains a predictive performance that is better than any
213
of the constituent models alone.
214

We evaluated two types of simple ensemble averages. The first ensemble was the
215
average of the 10-fold cross-validation, which was reported as the model performance.
216
This ensemble of 10 model weights was reported as one model because the architecture
217
and image exposure was the same. Only the training and validation data were different
218
in these models. The model weights were selected during training when the model had
219
the lowest MSE on the validation set. The average MSE and accuracy of the prediction
220
of the 515 test images from 10 folds on the test set were reported as the model MSE,
221
and accuracy.
222

The second ensemble was created from selections consisting of 2, 3, 4 models, and so
223
on up to an ensemble containing all 17 models. These ensembles combine 20, 30 and up
224
to 170 predictions on the test set. The accuracy was reported after rounding.
225

2.6 [Correlations of predictions on the test set and clustering analysis](#) 226

Correlations of predictions on the test set were investigated by creating a correlation
228
matrix of each model's prediction of each age class. This matrix showed how much the
229
models were in agreement, and clustering analysis identified which models were more in
230
agreement with each other. We used Pearson's correlation coefficient and hierarchical
231
clustering (HCA) with Euclidean distance and complete linkage.
232

2.7 [Comparison of CNN with human readers](#) 233

To evaluate the credibility of CNN predictions in relation to human readers, we
234
compared the mean percentage agreement of the test set predictions within each age
235

class with those from multiple human readers from a recent internal cod age reading
236 workshop carried out in 2021 at the Institute of Marine Research, Norway. In this
237 workshop, a ~~different~~ set of 100 broken otoliths from Atlantic cod were read by seven
238 readers, of which five were certified advanced cod readers and two were under training.
239 By comparing the results of the test set to the mean agreement and standard deviation
240 of predictions within the age class from the workshop, we evaluated if machine-driven
241 estimates were behaving in line with those anticipated by human readers.
242

Results

243

3 Results

244

The mean accuracy of the 17 models was 72.7% (Table 1) on the test-set, and the
245 standard deviation was 1.1. The least accurate model was B4-max with 70.9%, and the
246 most accurate model was B5-min and B6-middle with an accuracy of 74.4%.
247

B5 was the highest scoring model on all the exposures (min, middle, max) with a
248 mean accuracy of 73.7%, and min-exposure was the best exposure with a mean accuracy
249 of 73.3% Both B5 and B6 from the EfficientNet family were better than Medium and
250 Large from the EfficientNetV2 family.
251

The mean MSE of the 17 models was 0.284 on the test set, and the standard
252 deviation was 0.022. The highest MSE was from B5-max with MSE of 0.359, and the
253 lowest MSE was from B6-middle exposure with MSE of 0.262. We find that the
254 differences between models were statistically significant (ANOVA, $p = 1.6 * 10^{-7}$), but
255 that the differences were not significant for the individual factors of model architecture
256 (two-way ANOVA, $p = 0.139$) or image exposure ($p = 0.057$). ~~The interaction plot for~~
257 ~~two-way ANOVA is shown in Figure 2.~~ See Table 13 in the supplementary information
258 for a T-test of all models. From the interaction plot for two-way ANOVA, we see that
259 using low exposure images is more beneficial for the smaller models (Figure 2).
260

Medium and Large were the best models with a MSE of 0.278, and the 9-channel
261 composite images gave better results than any individual exposure, with a MSE of 0.272.
262 The high MSE for B5-max and B6-max was due to a large misprediction of the image
263 with index 308 in the test set labeled 1 year and predicted 5.7 years (see Table 7 in
264

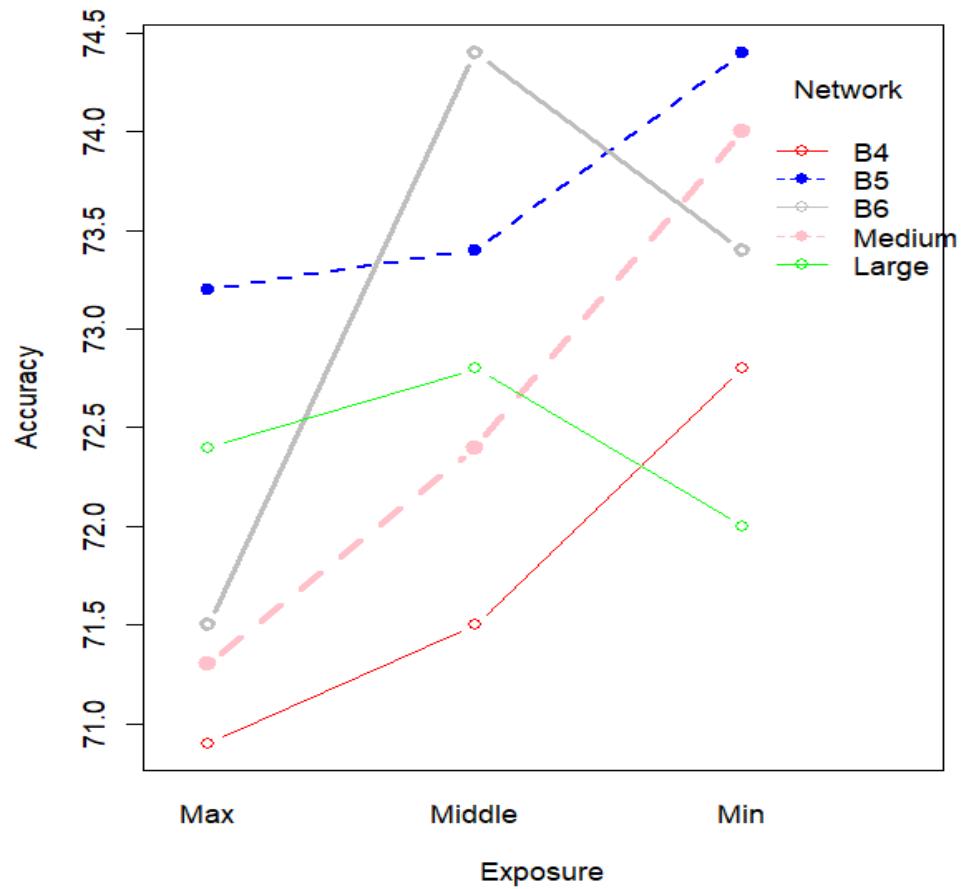


Figure 2. Interaction plot of the 5 networks with image exposure on x-axis and ensemble accuracy on the y-axis. We see that under-exposed images perform better for all but the B6 and Large network.

Table 1. Mean accuracy, MSE, and Percentage Agreement (PA) on the test-set by light exposure and CNN architectures

Acc:light/CNN	EfficientNet V1			EfficientNet V2		
	B4	B5	B6	Medium	Large	Mean
min	72.8	74.4	73.4	74.0	72.0	73.3
middle	71.5	73.4	74.4	72.4	72.8	72.9
max	70.9	73.2	71.5	71.3	72.4	71.9
9 channels	-	-	-	74.0	72.2	73.1
Mean	71.7	73.7	73.1	72.9	72.4	72.7
MSE:light/CNN						
min	.277	.277	.272	.273	.280	.276
middle	.285	.273	.262	.278	.275	.275
max	.291	.359	.305	.289	.286	.306
9 channels	-	-	-	.273	.271	.272
Mean	.284	.303	.280	.278	.278	.284
PA:light/CNN						
min	89.5	89.3	88.2	89.7	89.9	89.3
middle	88.2	89.5	90.9	91.1	87.8	89.5
max	87.6	90.5	88.0	89.5	90.3	89.2
9 channels	-	-	-	91.3	91.1	91.2
Mean	88.1	89.8	89.0	90.4	89.8	89.6

supplementary information on outliers).

Medium-all was the highest scoring model with percentage agreement (PA) 91.3%
and B4-max was the lowest scoring model with PA 87.6% (Table 1). Medium was the
overall best performing model and B4 was the worst. The 9-channel composite images
outperformed individual exposures, while max exposure had inferior performance to the
others.

When comparing each 10-fold ensemble average prediction accuracy, and MSE for all
17 models, the ensemble metric was either better than or in the upper quantile for all
the models (Figure 3). The prediction MSE and accuracy of each fold are given in
supplementary information (Table 5, 6).

3.1 Prediction by age class

When calculating the accuracy of all models by age class, we found that accuracy for
one- and two-year-olds was better than 90% (supplementary information in Figure 10).
All age classes six years or younger were correctly classified with more than 70%
accuracy, and all 13-year-olds were predicted to be younger.

No systematic bias in the age prediction of CNN is visible except for the
underestimated age of individuals aged by the expert reader as 13 years old (Figure 4).

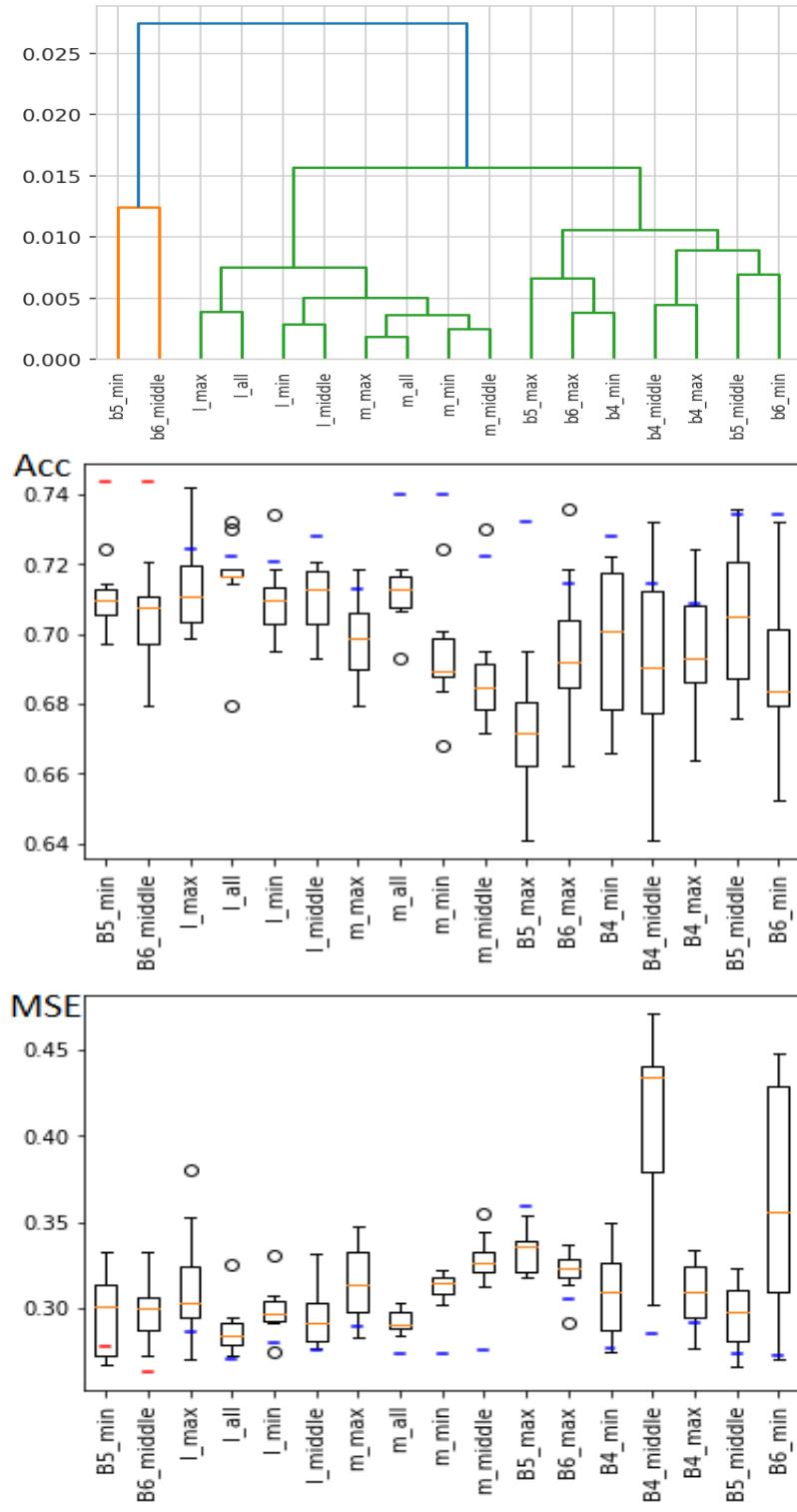


Figure 3. Hierarchical clustering (HCA) on the correlation of predictions (top), a box-plot of accuracy score (middle), and MSE (bottom) of all the 17 models. In (middle) and (bottom), the blue line is ensemble-average prediction accuracy (or MSE) on the test set, the red lines are the two best ensemble-average predictions on the accuracy, and the orange lines are the mean of the 10-fold predictions.

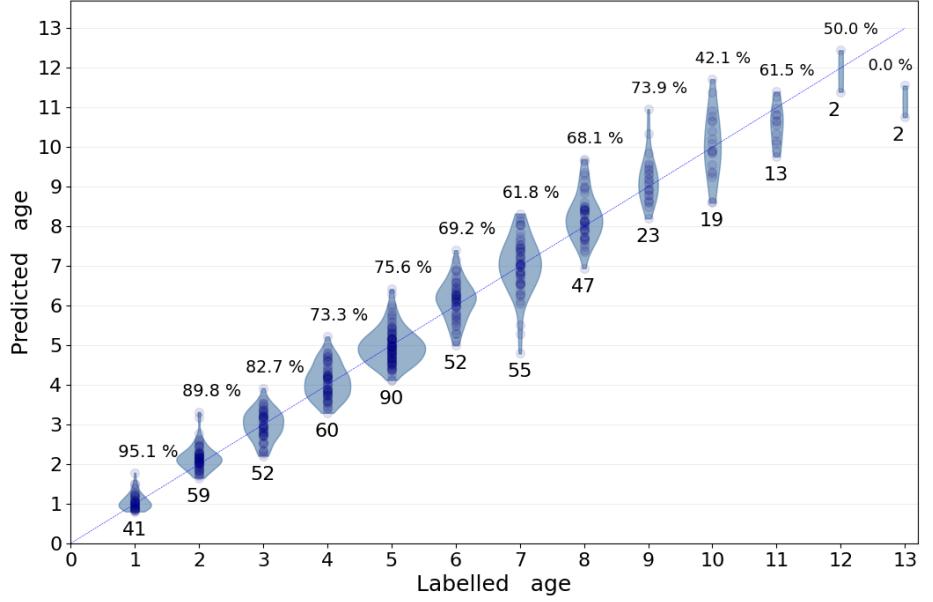


Figure 4. Violin plot of predicted age from model B5-min with accuracy of 74.4%. Above each age is the accuracy, and below is the total number of images in the test set of that age class

3.2 Simple ensemble-average predictions

We searched the space of ensembles-average predictions of 2 to 17 models, which is the set of unordered combinations without replacement, equal to the binomial coefficient $\sum_{k=1}^N \binom{N}{k}$ where $N = 17$ and $k \in 2..N$. For each set of ensemble combinations, we recorded the best ensemble and found that the best overall ensemble prediction was an ensemble of six models which produced an accuracy of 78.6%. The ensemble consisted of B4-min, B5-min, B6-min, Medium-min, B6-middle, and B4-max. The results are presented in [detail in](#) supplementary information in Table 10, 11, and 12.

The ensemble accuracy decreased after adding 6 models, while the MSE continued to decrease until all 17 models were included. This was as expected from the theory on simple ensemble average learning since the variance is reduced with more models.

We observe that ~~model-the models~~ B4-min (No 1), ~~and~~ B6-min (No 3) were the most often present in the top scoring model with inclusion in 14 ensembles (Table 2). These models did not have the highest accuracy (B5-min, and B6-middle) but an

Table 2. Rank statistics of models by participation in the best ensemble of size 1 to 17 when the loss function is accuracy.

Rank	Model name	Count
1	B4_min	15
1	B6_min	15
3	B5_min	13
3	M_min	13
5	B6_mid	12
6	B5_mid	10
6	B4_max	10
8	L_mid	9
9	B6_max	8
10	M_mid	7
10	M_all	7
10	L_all	7
13	M_max	6
14	L_min	5
14	B4_mid	5
14	B5_max	5
14	L_max	5

accuracy of 72.8% and 73.4%. This was lower than the highest accuracy models, which were B5-min and B6-middle (74.6%) with a rank of 3 and 5, respectively.

The mean ranking by exposure types was: min-exposure (rank 4.4), middle-exposure (rank 8.6), 9-channel composite (rank 10), and max-exposure (rank 11.2). The mean ranking by architecture was EfficientNet (rank 6.6), and EfficientNetV2 (rank 10.3).

3.3 Outliers

Figure 5 shows 4 images that were incorrectly classified with an error larger than 1 year after rounding. All the images with more than 1 year in prediction error are shown in supplementary information (Table 7), with comments by an expert on the most common mispredictions (Table 8). Large outliers occurred throughout all of the tested models and ensembles in small numbers. Most of those outliers were identified as visually challenging images with artifacts and/or low readability. For example, image 13 was overestimated in all B models, likely due to a clear zone in the inner core region that an expert reader would identify as a settlement false zone and ignore. Similarly, many outliers, such as images 270 and 369, showed multiple narrow false zones in the mid-section of the otolith that were likely to affect age determination. Alternatively, cases such as images 71 and 342 showed clear issues with age interpretation when the

image deviates from the standard of the training set, such as when the exposure was
314 changed drastically or when break lines interrupted the normal pattern of ring
315 deposition. In one case (image 362), all models estimated the otolith to be 5 instead of 7
316 years old: upon visual investigation, the otolith was clearly 5 years old, and the initial
317 age had likely been misread.
318



Figure 5. Example outlier images with index 13, 71, 279, 342 from the test-set were
mispredicted by between 25% and 100% of the models

Some cod otoliths were outliers to all models and on all exposures (e.g. otoliths 71,
319 342, 362, and 369), to a family of models and on all exposures (e.g. otoliths: 13, 423),
320 to some models and on one exposure (E.g otolith 308), and to both families of models
321 and on some exposures (E.g. otolith 320).
322

We also observed that the number of outliers did not correlate with model
323 performance. *E.g.*, B5-min, and B6-mid which had 7 and 9 outliers, but the best
324 accuracy. While B4-max with the lowest accuracy (70.9%) had the least number of
325 outliers with only 6 mispredictions.
326

3.4 Correlation of predictions and cluster analysis

The correlation of models on the test-set predictions given in Figure 11 in
328 supplementary information shows that the models strongly correlate in outlier
329 predictions. The correlation from all the predictions on the test set varied between
330

0.988 to 0.999, with the lowest correlation found between B5-min and Medium-min. 331

Hierarchical clustering (HCA) of the models found 3 clusters, ~~a One cluster~~ 332
~~contained~~ B5-min ~~, and~~ B6-middle, which ~~are the two best models, a cluster were the~~ 333
~~two best performing models. A second cluster contained~~ of all the EfficientNetV2 334
models, and a ~~cluster of~~ third cluster contained 335 the rest of the models (Figure 3, and
11). 336

The two least correlated models, B5-min and Medium-min, which had Pearson's 337
correlation of 0.988 showed strongly correlated predictions also on a sub-year scale 338
(Figure 6). This means that the model output is not simply a Gaussian or other 339
symmetric distribution around the correct (integral) value, but that the models identify 340
characteristics of the input that lead them to classify it to a highly precise fractional 341
value. 342

3.5 Comparison of CNN with human readers 343

Variations in percentage agreement between age classes showed similar patterns in both 344
CNN-based predictions and human readers, with generally decreasing agreement with 345
age (Figure 7). Within each age class, percentage agreement from CNN-based 346
predictions was lower than the average for multiple human readers and increasingly so 347
for the older age classes. However, they often remained within or close to the range of 348
percentage agreement observed across all readers for all otoliths of a given age class. 349

Discussion 350

4 Discussion 351

We successfully trained and tested machine learning methods on images of broken 352
otoliths, and achieved a maximum accuracy of 78.6% with an ensemble consisting of 353
six models, noting that the accuracy is the agreement between the read ages and the 354
model predictions. 355

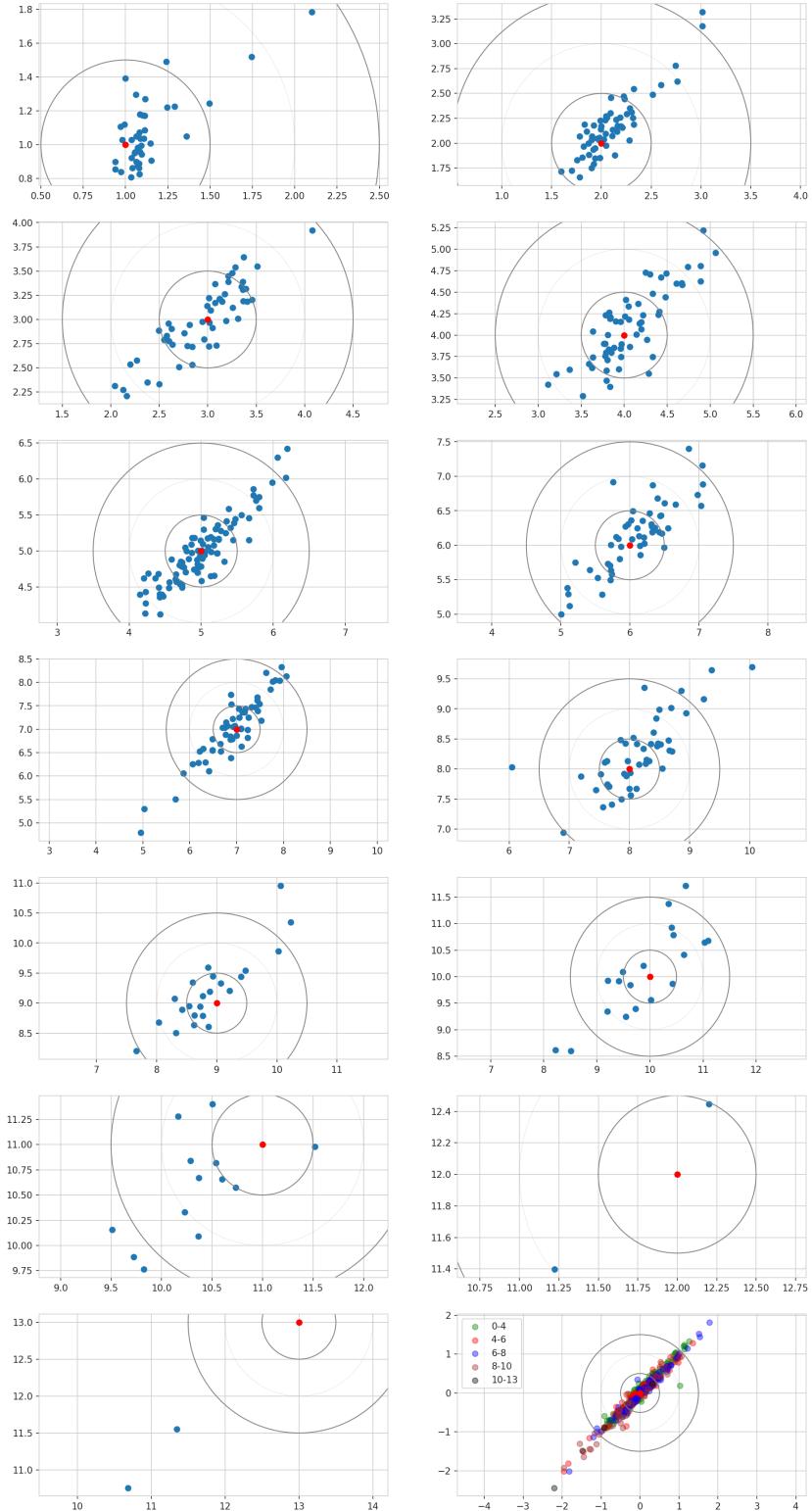


Figure 6. Comparison of age estimates predicted by Medium-min (x-axis, years) and B5-min (y axis, years) as age-specific scatter plots, and relative to in aggregate for all ages age groups in the bottom right corner panel. The x- and y-axis represents circles show age . The radius differences of the circle are r=0.5 and r=1.5 years.

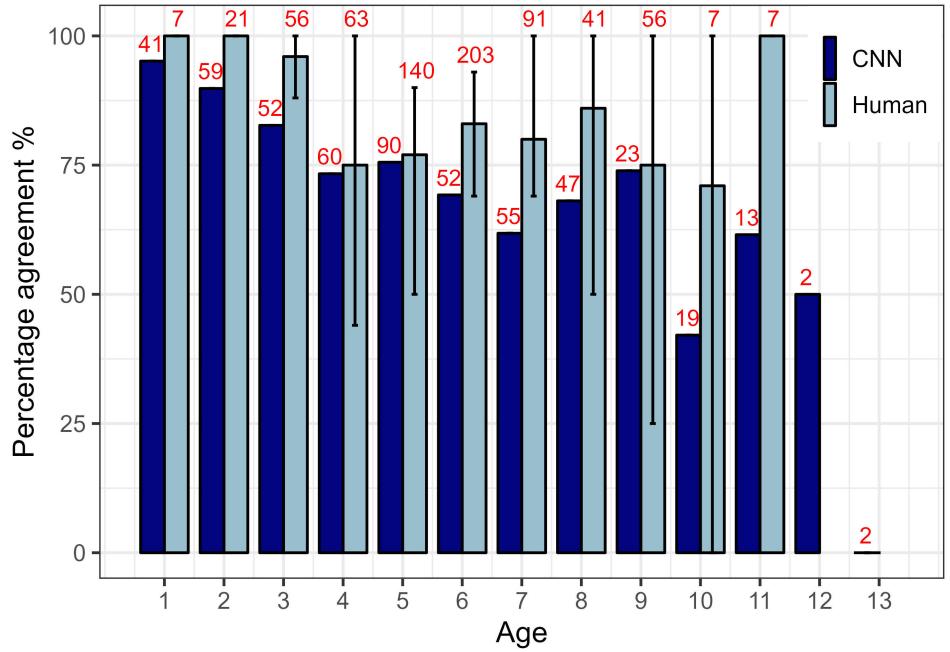


Figure 7. Comparison of mean percentage agreement within each age class for two sets of otoliths: the CNN-predictions on the test set (black); an internal age reading of 100 cod otoliths involving 7 readers (gray). Numbers indicate the total number of readings for each age class (with 1 reading per otolith for the CNN but 7 readings for the workshop). Error bars indicate the range of percentage agreement between readers for all otoliths of a given age class.

4.1 Accuracy across different age groups

The age of the younger individuals was predicted with greater accuracy than those of older individuals by our models. Thus, the CNN appears to be particularly competent at aging cod otoliths of younger age. This is also typical for expert readers who generally show the greatest accuracy for the youngest age classes which have fewer and clearer rings (Campana, 2001). However, the reasons both humans and CNNs find the age of younger individuals easier to predict may not be the same. Human expert readers use various visual cues, prior knowledge, and background information to determine fish age, such as comparing ring counts on multiple axes and having intrinsic knowledge of the periodicity of opaque and translucent zones for a given species. In younger fish, the increments are usually wider and more clearly separated as fish -and consequently otolith- growth rates are maximal prior to maturity. Fish of age 1 are small and have comparatively small otoliths with a straightforward ring pattern made of one single finished opaque and translucent zone, and expert readers are unlikely to disagree on it.

its interpretation. On the other hand, a CNN architecture as used here identifies 370
hierarchical patterns on different scales of the image from which it derives a value in the 371
range of those provided in the training set. This means that unless specifically forced to 372
do so, the algorithm may seek and interpret visual clues other than the rings human 373
readers are trained to use. A possible explanation for the higher prediction accuracy of 374
younger fish is that age is related to the area the otolith covers relative to the total 375
image size. Because the same camera settings were used, all images had the same 376
dimension and calibration. For a species with moderately large adults such as Atlantic 377
cod (Froese and Pauly, 2022), the otoliths will grow in size significantly faster during 378
the first years and then slow down with approaching sexual maturity. ~~This is especially 379
marked for the first couple of years where there is likely a more distinct size 380
stratification of the otoliths from one year to another.~~ As fish get older different 381
growth trajectories will then lead to greater overlap in otolith sizes across different ages. 382
It is therefore possible that the CNNs are not counting growth zones as human expert 383
readers would, but rather that they synthesize all available patterns in the image to find 384
recurring characteristics to the ages provided in the training set. The size of the area 385
that the otoliths cover against the more uniform black background might for example be 386
a very simple feature picked up by the CNNs yet with high predictive power for the 387
youngest fish, while the higher inter-individual variability and greater size overlap at 388
older ages would affect the predictive accuracy of CNNs 389

The hypothesis that CNNs exploit other information than the growth zones is 390
consistent with the findings of an earlier study where network activations inside a CNN 391
were explored for images of Greenland halibut otoliths (Ordonez et al., 2020). 392
Visualisation techniques were used to reveal the relative importance of attributes such 393
as shape, inner structure, and size of the otoliths using activation maps. Importantly, 394
the authors found that the CNN utilized information in pixels corresponding to annual 395
increments to only a small extent. ~~To explore this possibility, we attempted to train a 396
network using otolith silhouettes only, but were unable to achieve good performance.~~ 397

4.2 Importance of training set size relative to model

398

performance

399

It is commonly recognized that the performance of deep learning systems often improves with more training data (LeCun et al., 2015). A crucial issue in machine learning projects is then determining the amount of training data needed to achieve a specific performance goal. In this study we utilized a somewhat large data set of around 5000 images, although the images were divided among a large range of age classes. In comparison, it is not uncommon for deep learning systems for image classification such as ImageNet to be trained on thousands of images for each class (Russakovsky et al., 2014). In this study, the use of transfer learning (Yosinski et al., 2014) and augmentation yielded a significant performance boost but it is still likely that the network would provide more robust predictions with a larger training set. From a preliminary initial training not reported as part of this study, we trained a B4 network on about 2000 images and obtained an accuracy of around 60%. When another 3000 images were added to the data set accuracy reached about 70%. This could suggest that increasing our sample size would have further increased accuracy.

413

4.3 The effect of image size

414

The high-resolution 3744×5616 cod otolith images were scaled down to between 380×380 and 528×528 pixels to match the requirements of the different EfficientNet architectures. This reduction in resolution may have affected the readability of finer-scale visual features such as growth rings. The fixed camera setup resulted in the background constituting a non-negligible proportion of the images, especially for smaller otoliths. This is especially true due to the curved or oval shape of the otolith, as a compressed image will not only have less pixels to work with but will also have a comparatively more important fraction of black background, which is effectively useless for age interpretation. Otoliths of other fish species like red mullet which have a more circular shape may be less sensitive to this problem (Politikos et al., 2021). Improvements might therefore be made by instead first isolating the otoliths from their background and cropping the image accordingly, in order to have a machine learning network trained on using exclusively the information contained within the area of

424

425

426

427

interest. This would also limit information loss from image compression.

428

4.4 Outliers and transparency

429

The different networks were generally able to predict the age of otoliths with less than a
430 one-year deviation from the labelled values. It is noteworthy that predicted ages were
431 similar across different models and errors of more than a year were only seen in 2% of
432 the predictions on the test set. Closer inspection revealed that such errors were often
433 caused by otolith images with poor readability, in particular drastic changes in exposure
434 or visual damages and interruptions on the reading axis.
435

436

Interestingly, for one of these images, the predicted age was correct, and a
436 reexamination by an expert revealed that the initial annotation was wrong by two years.
437 While the previous results suggest that the network may not have relied entirely on ring
438 patterns for estimating age, this correct prediction of a wrongly assigned age shows that
439 it is still utilizing cues that are somewhat age-specific. Model behaviour was also similar
440 across all networks on single predictions of outliers: four were identified in all of the
441 models, suggesting they must have learned the same features.
442

443

4.5 Effects of image exposure on predictive power

443

Among the 17 models trained and explored in this study, models trained on
444 low-exposure images gave the best performance. Models trained on the
445 medium-exposure images and the nine-channel images also performed better than the
446 high-exposure images. The reason for this is not entirely clear ~~but has been observed in~~
447 ~~other studies~~. While low-exposure images may seem too dark and thereby hide useful
448 visual details from a human point of view, our results show that it is not necessarily the
449 case for an algorithm operating on finer-scale pixel values. It is likely that overexposure
450 causes burnout and irreversible loss of information while underexposed images retain
451 their information and only suffer from introduced noise.
452

453

4.6 Effects of 9-channel composite images and architecture size

453

Combining all 3 exposures into a 9-channel image did not perform better even if more
454 information was available to these CNNs. The EfficientNetV2 models trained on these
455

~~perform~~ performed similarly to models trained on single-exposure images. The variance 456
of predictions on 9-channel images was noticeably lower than for regular images, 457
meaning the CNNs were more certain in their prediction even when the prediction was 458
wrong. 459

We also found that the newer and larger EfficientNetV2 architecture did not stand 460
out as better than the EfficientNetV1 models. On the contrary, some of the best models 461
were the smaller ones of B4, B5, and B6. This could be due to the size of our data set 462
not being large enough to utilizing all the parameters in the larger models. Larger 463
networks are generally able to better explore a larger data set, such as ImageNet, 464
through training. 465

4.7 Utilising model ensembles.

We observed slight improvements in performance when an ensemble of models was used 466
for prediction. The use of numerous models in ensembles resulted in large numbers of 467
combinations of model predictions with varying accuracy. Some combinations achieved 468
higher accuracy than others (close to 79% for some combinations of six and seven 469
models). However, the mean ensemble prediction accuracy for a given number of models 470
showed that five models or more in combination resulted in accuracy just above 75%. 471
Five models thus seem to be sufficient and there could be minimal gain in precision in 472
combining larger numbers of models. Interestingly, ensembles combining models with 473
higher variance resulted in better predictions. This can indicate that if models are too 474
similar in individual predictions, the averaging effect ("wisdom of the crowds") will not 475
play out in the same way as when models with higher variance are combined. 476
Remarkably, many predictions only disagreed with a small decimal fraction. This could 477
imply that the models learned the same features in the otolith images. 478

4.8 Comparison of CNN with human readers

The comparison of age-specific percentage agreement in CNN-based predictions with 480
those from an internal age reading workshop ~~shows~~ showed that our models may 481
achieve similar agreement with human readers. While the numerical results are not 482
directly comparable in the sense that two different sets of otoliths were read, the trends 483
484

in mean precision across age groups were similar. Of particular interest is the fact that
485
the mean percentage agreement for our CNN-based predictions for a given age group
486
generally fell within or close to the range of percentage agreement for all otoliths of a
487
given age class seen between all readers involved in an age reading workshop. This may
488
indicate that while machine-based methods have not yet have the predictive accuracy of
489
an expert human reader, their estimates still fall within the expected range and may not
490
be easily distinguishable from those of traditional readers. Further testing should be
491
conducted to assess whether this is consistent, for example by conducting a multi-reader
492
aging event that includes undisclosed machine-based estimates of the same samples and
493
monitoring how they compare and whether they can be picked out by human readers.
494

4.9 Resource efficiency

Even if networks are reliable and trustworthy a remaining question will still be whether
495
there are significant cost benefits of deploying a ML framework for age reading of
496
otoliths. Despite fast progress, the results remain mixed and often yield lower precision
497
and consistency than those obtained by trained expert readers, which limits the
498
application of automated methods in real conditions. However, one aspect that is often
499
under-considered by such studies is the practical time and cost benefits that
500
implementing a functional ML framework would provide. As noted by [Fisher and](#)
501
[Hunter \(2018\)](#) ([Fisher and Hunter, 2018](#)) [Fisher and Hunter \(2018\)](#) in their review of
502
digital techniques for otolith analysis, “costs for human and machine ageing systems are
503
broadly similar since a large part of the cost is associated with preparing the otolith
504
sections”. As such, the net benefit of automated ageing routines is directly dependent on
505
the ability to scale performance using a comparatively smaller number of samples than
506
expert readers or, alternatively, to train them on “rougher” data that can be produced
507
faster and at a more efficient cost. Our study brings a net improvement toward this
508
resource-efficient inclusion of machine-driven analysis to age reading, as our networks
509
were trained exclusively on imaged broken otoliths. Whereas sectioned material requires
510
time and laboratory resources to embed, section, and prepare the samples for imaging,
511
breaking otoliths can be done immediately following collection from the fish. This
512
means that images and age estimates could potentially be produced directly at sea, or
513

at least processed in bulk as soon as the vessel and data are brought back to land. 515

Also, CNNs can be applied without high additional cost or even be incorporated into 516
the routine protocols, but add a new value e.g., reading consistency check, time-drift 517
evaluations, inter-reader comparisons (how much ‘off’ is each reader when compared to 518
the CNN predictions, even if not compared with the same otolith samples), etc. The 519
advantages of ensemble predictions will also be easier to gain with networks. Ensembles 520
of several expert readers are highly resource-demanding – especially when scaling to 521
huge datasets – while an ensemble of, say eight versus three networks only requires a 522
little bit more computation. 523

We see the process of CNN implementation as an evolution of the protocols, with an 524
intensive phase of model development and training. Through gradual improvement of 525
model reliability, CNNs could emerge as a complementary supportive tool for traditional 526
age estimations. The integration of those technologies could help scale the capacity of 527
age reading experts and improve the sampling of biological data and monitoring of 528
various fish stocks. 529

While the exact features used by the networks may differ from those interpreted by 530
human readers, one may be content with a trained network as a black box relying 531
entirely on its empirical accuracy. Deep learning techniques are particularly powerful in 532
detecting patterns in data (LeCun et al., 2015), and whether the networks actually 533
detect and count annual increments as the defining features or not, a causal relationship 534
between what the network attends to and the structure of the otoliths is likely. If the 535
network does not use growth zones as the primary features for prediction, this means 536
that it found useful correlating patterns unavailable or not obvious to the human eye. 537
This was demonstrated in the case of Greenland halibut otoliths where the shape of the 538
otoliths seemed to be the defining characteristic correlated with age (Ordonez et al., 539
2020). Machine learning frameworks may therefore be used as complementary age 540
readers to experts participating in otolith image interpretation workshops, as 541
informative input playing part of additional background information, or relied upon as 542
autonomous age readers without a subjective bias specific to increment identification 543
and counting. As efforts to develop machine-assisted frameworks increase, an important 544
question to ask is therefore whether they are to be trained at replicating human 545
methodology for specific tasks or instead given the freedom to generate new and often 546

not easily visualised methods.

547

4.10 Conclusion

548

Our results demonstrate that the use of deep learning techniques in the analysis of
549 otoliths have a major potential for facilitating automation. Standard model
550 architectures trained on sufficient training data specific to the use case can accurately
551 predict age from images of broken otoliths. We believe that carefully trained CNNs
552 could become a major component in automated pipelines that require minimal
553 processing and could be able to produce near at sea age estimates. In addition,
554 algorithmic age estimates could serve as a useful reference for training and evaluating
555 expert readers.

556

When developing the CNN framework for the automatic age estimation, ~~it is advised~~
557 ~~to include we found that the~~ B4 architectures ~~as they are were~~ quick to train, ~~and~~
558 ~~performs good and that they performed well.~~ Ensemble approaches can also be
559 ~~recommended if heavier considered if the increased~~ computing effort is not ~~detrimental~~
560 ~~a constraint~~ to the automation process, as they ~~could provide a can provide~~ more
561 robust and higher-performing ~~prediction predictions~~. For a quick-to-train ensemble, B5
562 and Medium ~~could might~~ be added. ~~These Our~~ results also indicate that the use of
563 slightly under-exposed images may be beneficial.

564

Acknowledgements

565

5 Acknowledgements

566

We thank Jane Godiksen and age readers and technicians from the Demersal Fish
567 research group for providing otolith age estimates and images used for this study. We
568 thank Erlend Langhelle for providing insight into the image-taking-protocol and age
569 interpretation of cod otoliths.

570

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S.,
Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine
learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467.* 571
- Albuquerque, C. Q., Lopes, L. C. S., Jaureguizar, A. J., and Condini, M. V. (2019).
The visual quality of annual growth increments in fish otoliths increases with latitude. 575
Fisheries Research, 220: 105351. 576
- Beamish, R. J. and McFarlane, G. A. (1995). A discussion of the importance of aging
errors, and an application to walleye pollock: the world's largest fishery. in recent 578
developments in fish otolith research. *University of South Carolina Press, Columbia,* 579
S.C., pp. 545–565. 580
- Brunel, T. and Piet, G. J. (2013). Is age structure a relevant criterion for the health of
fish stocks? *ICES Journal of Marine Science*, 70: 270–283. 582
- Campana, S. (2001). Accuracy, precision and quality control in age determination,
including a review of the use and abuse of age validation methods. *Journal of fish* 584
biology, 59(2):197–242. 585
- Carbonara, P. and Follesa, M. C. (2019). Handbook on fish age determination: a
mediterranean experience. *General Fisheries Commission for the Mediterranean.* 587
Studies and Reviews, 98: 1–179. 588
- Chollet, F. and others (2018). Keras 2.1.3. <https://github.com/fchollet/keras>. 589
- Denechaud, C., Smoliński, S., Geffen, A. J., Godiksen, J. A., and Campana, S. E.
(2020). A century of fish growth in relation to climate change, population dynamics 591
and exploitation. *Global Change Biology*, 26(10):5661–5678. 593
- Fisher, M. and Hunter, E. (2018). Digital imaging techniques in otolith data capture,
analysis and interpretation. *Marine Ecology Progress Series*, 598:213–231. 594
- Francis, R. C. and Campana, S. E. (2011). Inferring age from otolith measurements: a
review and a new approach. *Canadian Journal of Fisheries and Aquatic Sciences*. 596

<p>NRC Research Press Ottawa, Canada. https://cdnsciencepub.com/doi/abs/10.1139/f04-063 (Accessed 3 February 2022).</p> <p>Froese, R. and Pauly, D. (2022). Fishbase.</p> <p>Hidalgo, M., Rouyer, T., Molinero, J. C., Massutí, E., Moranta, J., Guijarro, B., , and Stenseth, N. C. (2011). Synergistic effects of fishing-induced demographic changes and climate variation on fish population dynamics. <i>Marine Ecology Progress Series</i>, 426: 1–12.</p> <p>Høie, H., Millner, R. S., McCully, S., Nedreaas, K. H., Pilling, G. M., and Skadal, J. (2009). Latitudinal differences in the timing of otolith growth: A comparison between the barents sea and southern north sea. <i>Fisheries Research</i>, 96: 319–322.</p> <p>Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, <i>Advances in Neural Information Processing Systems</i> 25, pages 1097–1105. Curran Associates, Inc.</p> <p>LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. <i>Nature</i>, 521(7553):436–444.</p> <p>Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. <i>neurips</i>.</p> <p>Mahé, K. (2009). Project no. 044132. <i>Automated FISH Ageing (AFISA): final activity report</i>.</p> <p>Mingxing Tan and, Q. V. L. (2021). Efficientnetv2: Smaller models and faster training. <i>CoRR</i>, abs/2104.00298.</p> <p>Moen, E., Handegard, N. O., Allken, V., Albert, O. T., Harbitz, A., and Malde, K. (2018). Automatic interpretation of otoliths using deep learning. <i>PLoS ONE</i>.</p> <p>Myers, S., Thorsen, A., Godiksen, J., Malde, K., and Handegard, N. (2019). An efficient protocol and data set for automated otolith image analysis. <i>GeoScience Data Journal</i>.</p>	<p>598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623</p>
--	---

- Ordonez, A., Eikvil, L., Salberg, A.-B., Harbitz, A., Murray, S. M., and Kampffmeyer, M. C. (2020). Explaining decisions of deep neural networks used for fish age prediction. *PLoS ONE*, 15(6):e0235013. 624
625
626
- Panfili, J., de Pontual, H., Troadec, H., and Wright, P. J. (2002). Manual of fish sclerochronology. <https://archimer.ifremer.fr/doc/00017/12801/> (Accessed 3 February 2022). 627
628
629
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc. 630
631
632
633
634
635
636
- Politikos, D. V., Petasis, G., Chatzispyrou, A., Mytilineou, C., and Anastasopoulou, A. (2021). Automating fish age estimation combining otolith images and deep learning: The role of multitask learning. *Fisheries Research*, 242:106033. 637
638
639
- Ragonese, S. (2018). Methuselah or butterfly? when fish age estimates and validations tell different stories. the case of the european hake (*merluccius merluccius* l. 1758) in the mediterranean sea. ””. 640
641
642
- Reglero, P. and Mosegaard, H. (2006). Onset of maturity and cohort composition at spawning of baltic sprat *sprattus sprattus* on the basis of otolith macrostructure analysis. *Journal of Fish Biology*, 68: 1091–1106. 643
644
645
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. 646
647
648
- Sigurdardóttir, A. R., Sverrisson, D., Jónsdóttir, A., Gudjónsdóttir, M., Elvarsson, B. D., and Einarsson, H. (2023). Otolith age determination with a simple computer vision based few-shot learning method. *Ecological Informatics*, 76:102046. 649
650
651

- Siskey, M. R., Wilberg, M. J., Allman, R. J., Barnett, B. K., and Secor, D. H. (2016). 652
Forty years of fishing: changes in age structure and stock mixing in northwestern 653
atlantic bluefin tuna (*thunnus thynnus*) associated with size-selective and long-term 654
exploitation. *ICES Journal of Marine Science*, 73: 2518–2528. 655
- Smoliński, S., Deplanque-Lasserre, J., Hjörleifsson, E., Geffen, A. J., Godiksen, J. A., 656
and Campana, S. E. (2020). Century-long cod otolith biochronology reveals 657
individual growth plasticity in response to temperature. *Scientific reports*, 10(1):1–13. 658
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the 659
inception architecture for computer vision. *CoRR*, abs/1512.00567. 660
- Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional 661
neural networks. *CoRR*, abs/1905.11946. 662
- Tyler, A. V., Beamish, R. J., and McFarlane, G. A. (1989). Implications of age 663
determination errors to yield estimates. *ICES Journal of Marine Science*, 108: 27–35. 664
- Vabø, R., Moen, E., Smoliński, S., Åse Husebø, Handegard, N. O., and Malde, K. 665
(2021). Automatic interpretation of salmon scales using deep learning. *Ecol. Inform.* 666
63, 101322 (2021). 667
- Wightman, R. (2019). Pytorch image models. 668
<https://github.com/rwightman/pytorch-image-models>. 669
- Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified 670
activations in convolutional network. *CoRR*, abs/1505.00853. 671
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features 672
in deep neural networks? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, 673
N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing* 674
Systems 27, pages 3320–3328. Curran Associates, Inc. 675

6 Supplementary information

676

5.1 Hyper-parameters

677

Supplementary information

678

Hyper-parameters

679

Table 3. Hyper-parameters on each model

Param/CNN	B4	B5	B6	Medium	Large
train_batch_size	8	8	16	8	8
img_size	380	456	528	384	384
val_img_size	380	456	528	384	384
steps_per_epoch	1600	1600	1600	1600	1600
epochs	150	150	250	450	450
early_stopping	-	-	-	40	40
early_stopping_patience	14	14	22	-	-
reduceLROnPlateau_patience	7	7	11	-	-

Medium all-, and min-exposures was run with steps_per_epoch=160

B6 has epochs=150, early_stopping_patience=14, and reduceLROnPlateau_patience=7

B4 min was run with img_size=456

Table 4. Hyper-parameters on all models, TensorFlow only (B4,B5, B6), and PyTorch only (Medium and Large)

Parameter	Value	TensorFlow	PyTorch
learning_rate	1e-05	v	v
n_fold	10	v	v
test_size	0.1	v	v
in_chans	3 or 9	v	v
reduceLROnPlateau_factor	0.2	v	x
which_exposure	min, medium, max	v	x
scheduler	CosineAnnealingLR	x	v
T_max	10	x	v
min_lr	1e-06	x	v
weight_decay	1e-06	x	v
which_exposure	min, medium, max, all	x	v

in_chans is the number of channels as input for the model. It was either 3 for an RGB image or 9 channels for 3 images.

5.1 Model architecture

680

Model architecture

681

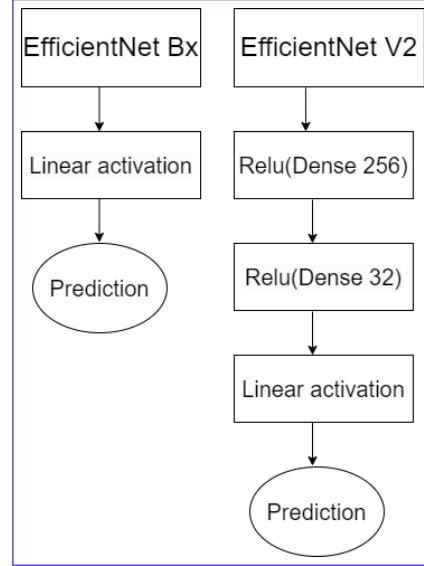


Figure 8. Diagram showing the changes made to the two architecture families

6 Descriptive Statistics

682

Descriptive Statistics

683

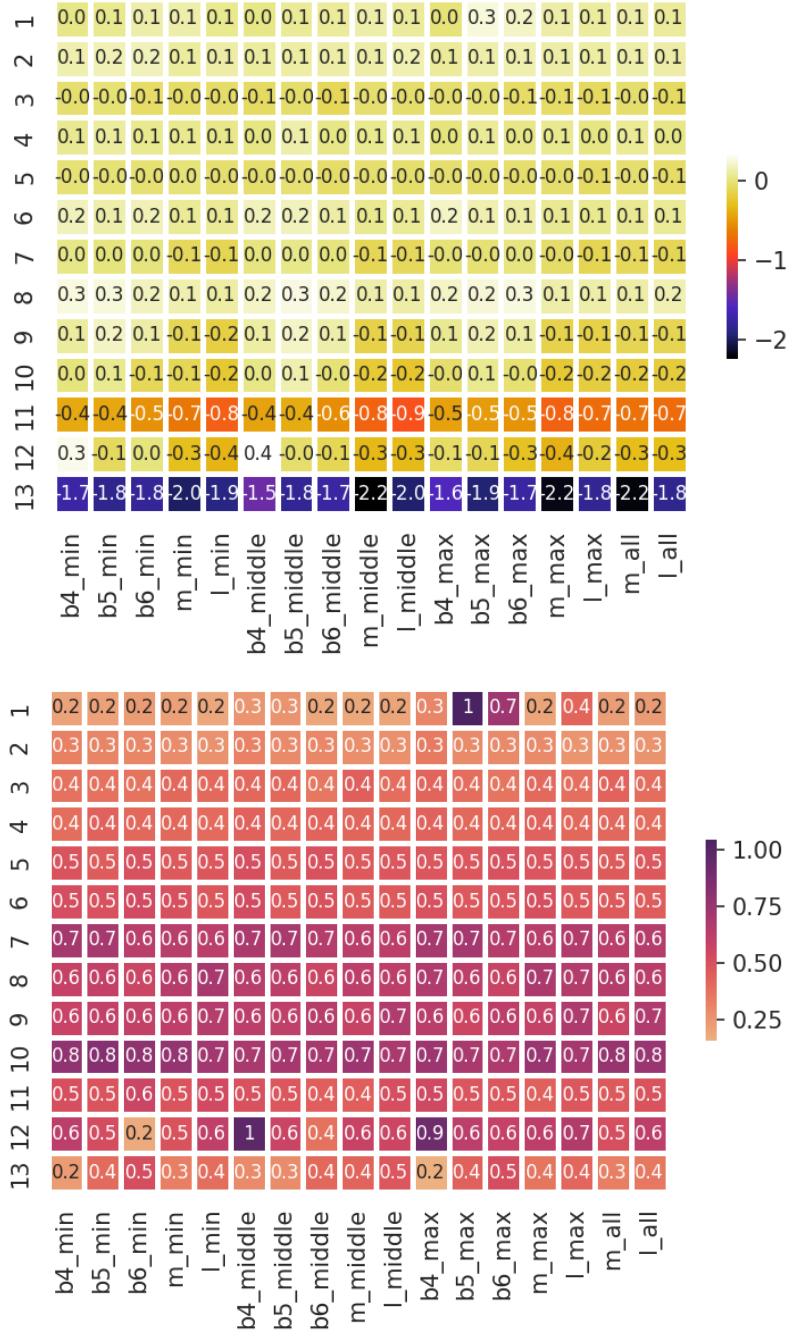


Figure 9. Model mean (top) and standard deviation (bottom) of residual test set prediction by age class

5.1 Model accuracy and MSE per fold

Model accuracy and MSE per fold

Table 5. MSE per CNN per fold

CNN/fold	1	2	3	4	5	6	7	8	9	10	ens.	Mean
B4,min	.320	.318	.306	.313	.322	.314	.315	.316	.306	.302	.277	.313
B4,middle	.344	.328	.316	.334	.326	.320	.355	.326	.313	.325	.285	.329
B4,max	.340	.317	.318	.347	.336	.336	.336	.320	.354	.336	.291	.334
B5,min	.324	.322	.325	.336	.291	.314	.320	.331	.33	.317	.277	.321
B5,middle	.308	.286	.315	.349	.332	.310	.280	.275	.331	.288	.273	.307
B5,max	.472	.302	.437	.459	.432	.366	.356	.441	.438	.418	.359	.412
B6,min	.325	.329	.334	.293	.312	.290	.320	.300	.276	.306	.272	.309
B6,middle	.323	.301	.312	.268	.294	.266	.309	.311	.278	.289	.262	.295
B6,max	.435	.306	.306	.270	.390	.321	.411	.321	.294	.448	.305	.350
m,min	.292	.292	.294	.275	.298	.304	.304	.331	.307	.295	.273	.299
m,middle	.287	.302	.307	.332	.288	.276	.277	.294	.304	.278	.278	.295
m,max	.337	.297	.302	.291	.315	.347	.338	.321	.313	.283	.289	.314
m,all	.289	.299	.303	.284	.292	.287	.303	.288	.289	.294	.273	.293
l,min	.267	.316	.269	.270	.322	.332	.280	.307	.303	.299	.280	.297
l,middle	.300	.332	.320	.300	.272	.302	.294	.285	.307	.285	.275	.300
l,max	.322	.295	.324	.353	.295	.306	.271	.292	.380	.299	.286	.314
l,all	.285	.293	.283	.274	.286	.325	.272	.283	.277	.295	.271	.287
Mean	.328	.308	.316	.315	.318	.313	.314	.314	.318	.315	.284	.316

686

Table 6. Accuracy per CNN per fold

CNN/fold	1	2	3	4	5	6	7	8	9	10	ens.	Mean
B4, min	69.9	68.9	68.7	68.3	68.9	70.1	69.7	66.8	68.9	72.4	72.8	69.3
B4, middle	68.5	69.3	73.0	68.5	67.8	68.2	67.2	67.2	68.3	69.5	71.5	68.8
B4, max	64.1	68.2	67.2	66.2	67.8	69.5	67.2	69.3	66.2	65.2	70.9	67.1
B5, min	71.8	69.1	69.3	66.8	73.6	70.7	66.2	68.3	69.5	68.7	74.4	69.4
B5, middle	70.3	72.0	67.8	66.6	67.4	69.9	71.8	71.5	68.2	72.2	73.4	69.8
B5, max	71.3	71.1	67.4	73.2	66.4	68.9	64.1	69.1	68.7	71.8	73.2	69.2
B6, min	68.3	68.5	66.4	72.4	70.7	70.9	69.3	69.3	72.0	68.9	73.4	69.7
B6, middle	68.5	69.9	67.6	73.6	72.8	72.0	68.0	69.3	72.0	71.1	74.4	70.5
B6, max	70.5	68.2	65.2	73.2	69.1	67.8	68.0	68.0	72.8	68.5	71.5	69.1
m, min	71.1	71.1	69.5	73.4	71.8	70.9	70.9	69.7	70.1	71.5	74.0	71.0
m, middle	71.3	70.1	70.1	70.9	71.7	71.8	72.0	71.3	69.3	71.8	72.4	71.0
m, max	68.9	70.1	70.3	71.3	70.7	68.5	69.7	68.0	69.1	71.8	71.3	69.8
m, all	71.7	70.7	69.3	71.3	71.8	71.8	71.3	71.7	71.1	70.7	74.0	71.1
l, min	72.4	69.7	71.5	70.8	71.3	71.3	70.9	69.9	71.1	70.5	72.0	71.0
l, middle	68.7	68.0	69.7	71.8	71.1	71.1	69.7	70.5	71.1	72.0	72.8	70.4
l, max	71.1	70.1	69.9	74.2	72.8	71.1	72.2	71.1	71.1	70.1	72.4	71.4
l, all	71.8	71.7	71.8	71.7	71.7	68.0	73.2	71.7	73.0	71.5	72.2	71.6
Mean	70.0	69.8	69.1	70.8	70.4	70.1	69.5	69.6	70.1	70.5	72.7	70.0

687

Predicted age class for all models and ground truth

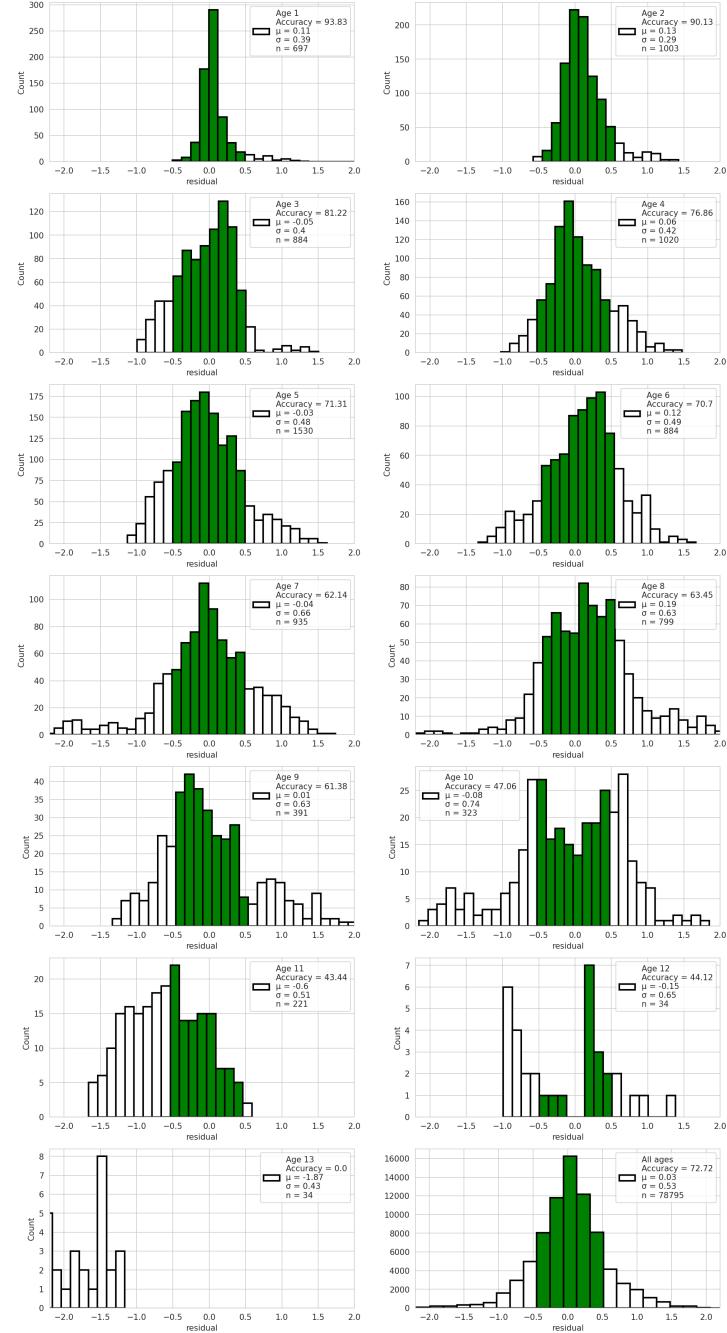


Figure 10. Predictions by age class from the average of all models. The green region shows the correctly classified age after rounding. The axis is fixed, hence outliers that differ from the true age by more than two years will not be visible.

6 Outliers

690

Outliers

691

Table 7. Predictions error with residual of more than 1.5 years per model per index in test-set

Idx	13	17	47	48	71	92	154	270	279	308	312	320	334	342	362	369	393	418	423	444	462	481	502	Count
B4-min	9.8				5.1		11.7	9.9		5.5		11.1	5.1	8.2									8	
B4-mid	9.7				5.4		10.2			5.4	7.5	11.3	4.9	8.3	10.6	9.5							10	
B4-max	9.6				5.0		10.4					11.3	5.0	8.2									6	
B5-min	9.6				4.8		11.7	9.7				10.8	5.3										7	
B5-mid	9.8				6.7	11.5	11.8	9.8				10.9	5.3	8.4									9	
B5-max	9.8				4.5	11.5	9.6	7.7				10.6	5.1	8.3									8*	
B6-min	9.7				7.6	5.1	9.7					10.7	5.2	7.9	10.8	10.7							9	
B6-mid	9.6				5.1		11.5	9.7				10.8	5.2	8.3	10.8								9	
B6-max	9.8				5.2				5.7			10.7	5.2	8.2	10.6								9**	
m-min					5.0	11.3		10.0				10.7	5.0	8.2		6.0							7	
m-mid					4.9	11.2		10.0				10.3	5.1	8.2									6	
m-max					6.5	5.1	11.2	8.7	10.2			10.5	5.1	8.1		6.3							9	
m-all					5.0	11.2		10.1				10.5	5.3	8.2		6.2							8	
l-min					5.1	11.5		9.8	9.3			10.7	5.2	8.3		5.1							8	
l-mid					5.0		9.8		9.4	5.5		10.6	5.2	8.1	10.5	6.0							9	
l-max					5.1			9.9	3.6	5.4		10.8	5.1	8.2		5.9							10	
l-all					5.0			9.8				10.8	5.2	8.0	10.5	6.2							9	
Age	8	8	8	6	7	13	7	10	8	1	11	7	6	13	7	10	9	11	8	11	5	10	-	
Count	9	2	1	1	17	7	1	4	16	3	2	2	1	17	17	6	2	7	2	1	3	3	141	
As pct	5.3	12	6	6	100	41	6	24	94	18	12	12	6	100	100	35	12	41	12	6	18	18	-	

692

Table 8. Comments on the most frequently mispredicted otolith images

Idx	Comment
13	Labeled 8 years, and read as 10 years by the B-models (EfficientNet). The quality of the exposures was good, but there was a lot of split rings in the middle.
71	Labeled 7 years, and read as 5 years by all models. The exposures was-were very bright on all three axis <ins>axes</ins> , and the dorsal axis had a break line (physical break or fissure <ins>fissure or physical break</ins>), and the plane was out of focus.
279	Labeled as 8 years, and read as 10 years by almost all models except B6-max. The exposures was-were of good quality, but there was-were split rings in the middle.
308	Labeled as 1 year, and read as 8 years, 6 years and 4 years by B5-max, b6-max, and Large-max respectively. The exposures was-were of good quality and the predicted age is obviously wrong.
342	Labeled as 13 years, and read as 11 years by all models. The quality of the exposures was good. The inner section is dark on the ventral side, the distal side is light, and the dorsal side has a break line. <small>693</small>
362	Labeled as 7 years, and read as 5 years by all models. This image is mislabeled. The otolith is obviously 5 years old.
369	Labeled as 10 years, and read as 8 years by all models except B5-min. The quality of the exposures was good, but it had split rings in the middle on bright exposures, and the contrast is strong.
393	Labeled as 9 years, and was read as 11 years by B4-middle, all B6 exposures and Large-middle and -all. The middle and min exposures was-were too dark. Max exposure was nice.
423	Labeled as 8 years, and read as 6 years by all the EfficientNetV2 models except Medium-middle. The quality of the images was bad. All the exposures was-were over-exposed.

694

6 Ensembles by simple average

695

Ensembles by simple average

696

Table 10 shows the number of combinations of models that exist of tuples, triplets, and so on labeled with the heading "Coeff", then the best ensemble-average accuracy on the given number of combinations, and then the model numbers that produced the best combinations. Model number can be translated to model name using Table 9. Table 11 shows the same information but selected to minimize MSE.

697

698

699

700

701

Table 9. The table shows the model family as columns and image exposure as rows. The numbering of models is used in reference to ensembles.

CNN family / Image exposure	EfficientNet			EfficientNetV2	
	B4	B5	B6	Medium	Large
Minimum	1	2	3	4	5
Medium	6	7	8	9	10
Maximum	11	12	13	14	15
9 channels	-	-	-	16	17

Table 10. Binomial combinations of simple average of ensembles accuracy

Coeff	#Comb	Best	Mean	Ensemble (see table 9)
2	136	75.9	74.1	(2, 5)
3	680	77.5	74.6	(1, 3, 4)
4	2380	77.9	74.9	(1, 2, 3, 4)
5	6188	77.9	75.1	(1, 2, 3, 4, 11)
6	12376	78.6	75.2	(1, 2, 3, 4, 8, 11)
7	19448	78.1	75.2	(1, 2, 3, 4, 7, 8, 11)
8	24310	77.5	75.2	(1, 2, 3, 4, 7, 8, 10, 11)
9	24310	77.5	75.3	(1, 2, 3, 6, 7, 8, 9, 11, 17)
10	19448	77.1	75.2	(1, 2, 3, 6, 7, 8, 9, 10, 12, 13)
11	12376	76.9	75.2	(1, 2, 3, 4, 6, 7, 8, 10, 11, 13, 16)
12	6188	76.7	75.2	(1, 3, 4, 7, 8, 10, 11, 13, 14, 15, 16, 17)
13	2380	76.3	75.1	(1, 3, 4, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17)
14	680	75.9	75.1	(1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 16, 17)
15	136	75.7	75.0	(1, 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17)
16	17	75.5	75.0	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17)
17	1	74.8	74.8	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)

702

703

704

Table 11. Binomial combinations of simple average of ensembles MSE

Coeff	#comb	best	Mean	Ensemble (see table 9)
2	136	0.250	0.265	(3, 17)
3	680	0.246	0.259	(1, 3, 5)
4	2380	0.245	0.256	(1, 3, 5, 7)
5	6188	0.245	0.254	(1, 3, 4, 7, 17)
6	12376	0.244	0.252	(1, 2, 3, 5, 8, 16)
7	19448	0.244	0.251	(1, 2, 3, 4, 5, 8, 11)
8	24310	0.244	0.251	(1, 2, 3, 4, 5, 8, 11, 17)
9	24310	0.244	0.250	(1, 2, 3, 4, 5, 7, 8, 11, 17)
10	19448	0.244	0.250	(1, 2, 3, 4, 5, 7, 8, 11, 16, 17)
11	12376	0.245	0.250	(1, 2, 3, 4, 5, 7, 8, 10, 11, 13, 16)
12	6188	0.245	0.249	(1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 16, 17)
13	2380	0.245	0.249	(1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 16, 17)
14	680	0.245	0.249	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 16, 17)
15	136	0.246	0.249	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 16, 17)
16	17	0.247	0.248	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17)
17	1	0.248	0.248	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)

Table 12. Comparison of the mean of all the 17 models (mean) with a total accuracy of 72.7% and the best ensemble model (Best Ens.) with a total accuracy of 78.6%. In all age groups, the ensemble improves on the mean-model accuracy except 13 year-olds.

Age	1	2	3	4	5	6	7	8	9	10	11	12	13
Mean	93.8	90.1	81.2	76.9	71.3	70.7	62.1	63.5	61.4	47.1	43.4	44.1	0
Best Ens.	95.1	93.2	84.6	80.0	78.9	78.9	65.6	76.6	69.6	52.6	61.5	50.0	0

5.1 Pearson correlation of each model on test-set predictions

705

Pearson correlation of each model on test-set predictions

706

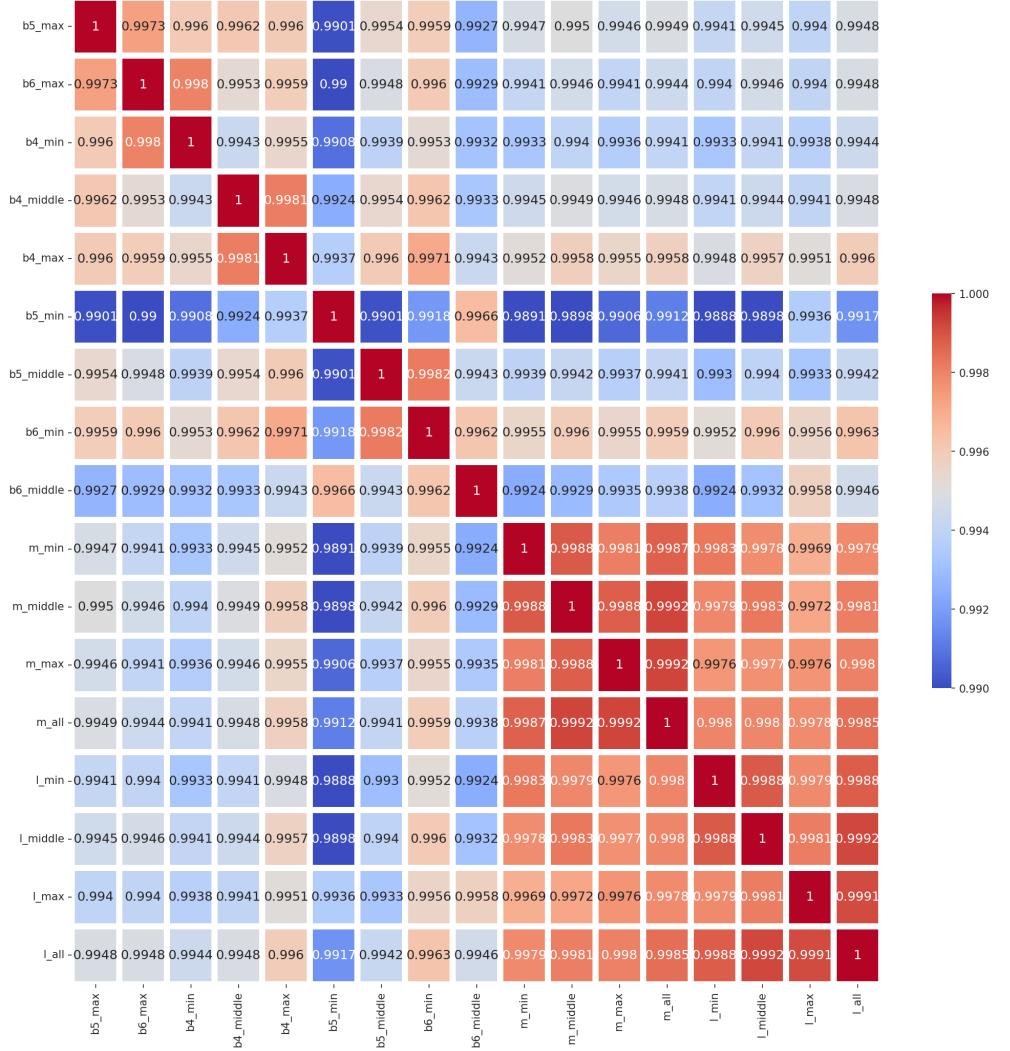


Figure 11. Pearson correlation of each model prediction on the test-set. The ~~gradient is from colors indicate the least correlated model of correlation coefficient, ranging from 0.99 (blue) to the most of 1.00 (red).~~

5.1 T-statistics of each model vs model comparison on test-set

707

~~prediction~~

708

~~T-statistics of each model vs model comparison on test-set~~

709

~~prediction~~

710

Table 13. T-statistics of each model vs other models (order as in table 1)

no	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
1	0.873	0.588	0.00901	0.0074	0.465	0.56	0.158	0.00554	0.0967	0.00645	0.938	0.864	0.361	0.00385	0.00322	0.00199	
2	-	0.768	0.0651	0.0679	0.462	0.723	0.284	0.0541	0.261	0.0173	0.851	0.782	0.601	0.0314	0.0407	0.0188	
3	-	-	0.0748	0.0755	0.251	0.931	0.381	0.0584	0.355	0.00417	0.649	0.566	0.826	0.0333	0.0408	0.0186	
4	-	-	-	-	0.86	0.00312	0.132	0.517	0.933	0.273	2.04E-05	0.0823	0.0467	0.0433	0.516	0.756	0.306
5	-	-	-	-	-	0.00293	0.139	0.562	0.764	0.282	2.65E-05	0.0875	0.0491	0.0348	0.381	0.546	0.205
6	-	-	-	-	-	0.256	0.0597	0.00219	0.0293	0.0414	0.6668	0.698	0.118	0.00138	0.00147	0.000743	
7	-	-	-	-	-	-	0.46	0.111	0.46	0.00652	0.616	0.539	0.922	0.0646	0.084	0.0389	
8	-	-	-	-	-	-	-	0.471	0.885	0.00104	0.26	0.199	0.422	0.289	0.388	0.189	
9	-	-	-	-	-	-	-	-	0.213	1.78E-05	0.0725	0.0396	0.0259	0.532	0.8	0.306	
10	-	-	-	-	-	-	-	-	-	0.000173	0.25	0.177	0.371	0.115	0.139	0.0611	
11	-	-	-	-	-	-	-	-	-	-	0.057	0.0453	0.0007539	0.82E-06	1.58E-05	5.84E-06	
12	-	-	-	-	-	-	-	-	-	-	-	0.947	0.514	0.0447	0.0586	0.0294	
13	-	-	-	-	-	-	-	-	-	-	-	-	0.418	0.0234	0.0307	0.0145	
14	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0174	0.0134	0.00865	
15	-	-	-	-	-	-	-	-	-	-	-	-	-	0.643	0.711	-	
16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.375	