

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 composition of fish populations plays a crucial role in stock management and provides valuable information for biological studies. Fish age is typically estimated by manually counting experts that manually count annual increments in otoliths, but this process is prone to age reader bias and, which makes comparisons between readers and labs challenging, 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 age estimation of otolith images using pictures of fish age prediction using 5150 broken images of otoliths from Barents Sea Atlantic cod (*Gadus morhua*). We used collected between 2012 and 2018. In contrast to previous studies that utilise models trained on otolith sections, we used images of broken otoliths that require no processing prior to imaging, and hence, could potentially facilitate at-sea age estimation. We trained convolutional neural networks (CNNs) based on two modern architectures (EfficientNetV1 and EfficientNetV2 with three and two different model sizes, respectively, from each type and compared the performance. The average accuracy and mean squared error when comparing predictions and the manually read ages for the tested models), which vary in model

size (number of model parameters), and compared performance. Model average
accuracy was 72.7% and mean-squared-error was 0.284, respectively when compared
with the human-read ages. The models' accuracy for one and two year old one- and
two-year-old individuals was over 90% and no systematic bias in the age predictions
across age groups was observed. The best models were EfficientNet B4 and
EfficientNet B6 using images taken with low exposure times. After an exhaustive
search, a maximum accuracy of 78.6% was achieved with using an ensemble
consisting of six models. The tested models strongly correlate in terms of
predictions. Model predictions were also strongly correlated, limiting the utility of
building large ensembles. Model performance was compared to the results of an
internal workshop where 100 independent images of broken otoliths were aged by a
group of experts. Variations in percentage agreement between age classes showed
similar patterns. A similar pattern (decreasing with age) in both CNN-based predictions
and human readers, with generally decreasing agreement with age. While percentage
agreement from CNN-based predictions age estimates made by the expert group.
While CNN-based percentage agreement was often lower than for human experts the
expert estimates, it remained within or close to the range of percentage agreement
observed across all readers. Our results demonstrate the potential of deep learning
techniques for analyzing otoliths extracting age estimates from otolith images. When
developing frameworks for age estimation using machine learning, we recommend
EfficientNet B4 models are used as they are quick to train quicker to train than larger
models and perform well. Ensemble approaches are also recommended if sufficient
computational resources are available, as they can give provide increased accuracy and
lower variance of the predictions. We find that under-exposed images tend to perform
better than 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.

1 Introduction

Knowledge of fish age structure is central to fisheries science and stock dynamics modeling. It informs modelling stock dynamics. It provides information 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., 2011; Brunel and Piet, 2013). Changes Monitoring changes in the age distribution can track significant of a fish population can help to track substantial changes in population structure, such as the appearance of a particularly strong year-class (Reglero and Mosegaard, 2006), or the gradual truncation of older age classes as selective fishing mortality removes larger individuals (Siskey et al., 2016). Hard structures such as scales and otoliths are used worldwide as one of the primary sources of fish age estimates, due to their ability as natural physiological and environmental recorders to form regular, which form regular and temporally resolved growth increments at the can be used to estimate fish age at an annual level (Campana, 2001; Francis and Campana, 2011; Albuquerque et al., 2019). While age is inferred from the “simple” counting of annual increments, the interpretation of this zonation pattern is species or even population-specific (Høie et al., 2009) and is based on precise knowledge of the timing of zone formation and of the correct identification of true and false zones (Panfili et al., 2002). This process, therefore, requires specific expertise and is subject to uncertainties in both between-reader precision and “true” age accuracy (Francis and Campana, 2011). These biases can occur between readers both within and between otolith laboratories. Therefore, streamlining, scaling, and increasing the quality of age estimations can improve the reliability of evaluations of fish biology and consequently assessment of stock size and structure (Tyler et al., 1989; Beamish and McFarlane, 1995; Ragonese, 2018).

Otolith reading is time and resource-consuming. Training of expert readers can take several years depending on the species, and otoliths often undergo a long processing phase before the final age estimates can be produced (Carbonara and Follesa, 2019). This is particularly true for demersal fish species, like Atlantic cod (*Gadus morhua*), that have large opaque and relatively untransparent otoliths that typically require time-consuming preparation (Denechaud et al., 2020;

Smoliński et al., 2020). These routines vary between species and populations and
77 institutes and range from a direct reading of broken otoliths under a magnifying glass,
78 to embedding, thin sectioning, and finally imaging of the sections under a microscope.
79 All otoliths read in Norway and Russia for the Northeast Arctic cod population are read
80 on broken otoliths using a magnifying glass.
81

There has been a variety of methods proposed to automatically interpret otoliths
82 from images, which range from one-dimensional data analysis like intensity transects
83 (Mahé, 2009) to the more recent effort toward developing machine learning (ML)
84 frameworks (Moen et al., 2018; Politikos et al., 2021; Sigurdardóttir et al., 2023). One
85 of the main advantages of automation is that the results are reproducible and consistent.
86 Age predictions obtained from the automatic algorithm can be used e.g. for the quality
87 control and identification of age reader biases within and between otolith laboratories
88 (ICES, 2013)
89

1.1 Deep learning and image analysis

During the last decade, deep learning has become one of the ~~dominating fields of~~
91 ~~dominant fields in~~ machine learning where various architectures of deep neural networks
92 are ~~able to learn to trained and used to~~ efficiently identify patterns and structures in
93 various types of data (LeCun et al., 2015). Within the field of computer vision, deep
94 Convolutional Neural Networks (CNN) have been ~~widespread commonly used~~ ever since
95 Krizhevsky et al. (2012) won the annual ImageNet Large Scale Visual Recognition
96 Challenge (ILSVRC) competition (Russakovsky et al., 2014). ILSVRC remains the most
97 important benchmark for image classification with 1.4 million images in the ImageNet
98 training set, and state-of-the-art CNNs are therefore often ~~targeted optimized for~~ to
99 this data set. Many of these CNNs ~~are publicly available including (and~~ their trained
100 network weights. ~~It is therefore often useful to use transfer learning with these) are~~
101 ~~publicly available, and are often used as the starting point for new CNNs, a process~~
102 ~~which is known as ‘transfer learning’. Using pre-trained weights as a starting point.~~
103 ~~This is especially true for tasks where relatively networks in this manner, can be~~
104 ~~particular effective when~~ little training data is available. For many fish species, age
105 estimation from images of otoliths represents precisely such a task. InceptionV3
106

(Szegedy et al., 2015) was modified to predict the age of Greenland halibut (107
Reinhardtius hippoglossoides) from otolith images (Moen et al., 2018), and a modified (108
InceptionV3 was applied to classify otolith images of red mullet (*Mullus barbatus*) (109
(Politikos et al., 2021). While some state-of-the-art CNNs ~~grew have increased~~ in model (110
size (i.e. number of model parameters) over time, a recent CNN architecture called (111
EfficientNet (Tan and Le, 2019) demonstrated that increased performance could be (112
achieved with smaller model sizes (number of parameters) using a compound scaling (113
method for network depth, width and image size, resulting in a family of seven different (114
models with different sizes. This network has been successfully applied with transfer (115
learning to analyse images of salmon scales (Vabø et al., 2021). Recently, a successor to (116
the EfficientNet architecture, EfficientNetV2 (Mingxing Tan and, 2021), has been made (117
available. (118

The main objective of this study ~~is was~~ to develop a deep learning framework for (119
automating the age estimation of Atlantic cod based on images of broken otoliths taken (120
with constant illumination and three different exposures. We tested EfficientNetV1 and (121
EfficientNetV2 ~~architectures~~ ~~architecture families~~ using a range of model sizes from (122
each~~family~~, and we compared the performance of ~~the different models, including as~~ (123
~~ensembles. We aim to individual models and ensemble model runs. We also~~ provide (124
best practices and strategies for developing CNN frameworks for ~~the predictions of the~~ (125
~~age of fish based on fish age predictions based on the images of~~ otoliths. We further (126
anticipate that this can serve as a baseline for the future development and (127
operationalization of CNN models and the inclusion of ML-based otolith age (128
interpretations ~~in the biological data collection routines during research expeditions~~ (129
~~and stock assessment surveys.~~ (130

2 Method and materials (131)

2.1 Data Collection (132)

We used a data set sampled from 5150 cod otoliths collected on surveys conducted by (133
the Institute of Marine Research (IMR) in the period 2012-2018 and aged by expert cod (134
readers. On each of the surveys, the otoliths were sampled using a random-stratified (135

sampling based on fish length for each trawl station.

136

The otoliths were sampled over a wide range of ages (1-13 years) but did not include
137 age-0 fish. Each otolith was broken in the transverse plane and placed on a mount
138 before it was captured by six images with three light exposures and one rotation of 180°
139 (Figure 1). The images were taken with a resolution of 3744×5616 pixels. The image
140 light exposure punctually varied depending on light conditions coming from outside.
141 Light exposure was stored in the metadata of the JPG file. Details can be found in
142 (Myers et al., 2019) and in the data set available at
143 <https://doi.org/10.21335/NMDC-1826273218>.

144

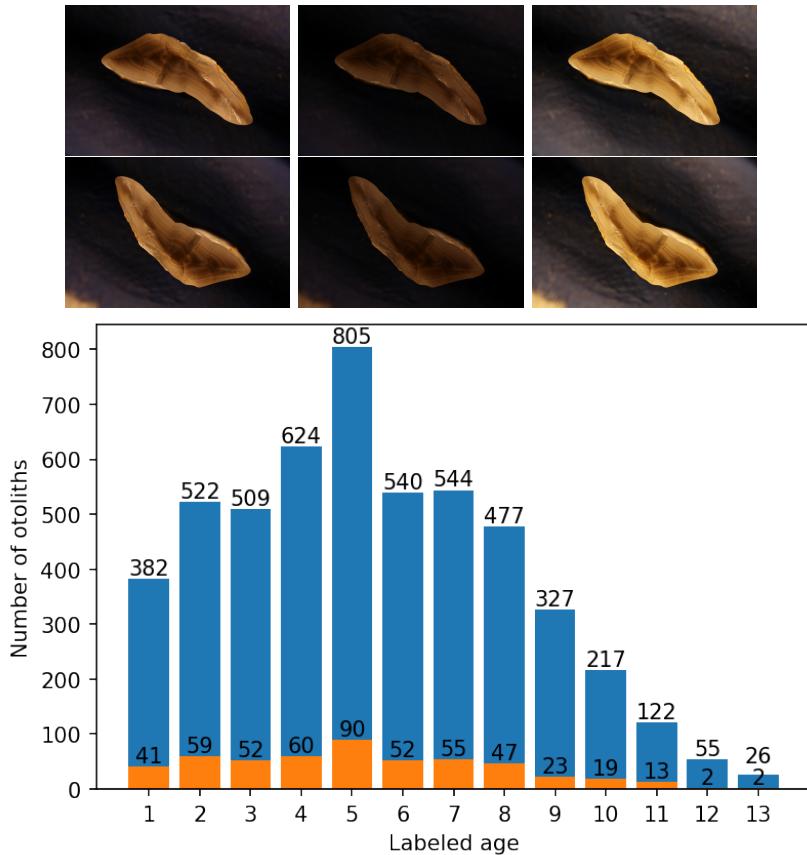


Figure 1. Images of an otolith collected in 2016 from a 6 years old cod (top), taken with medium-, minimum- and max-exposures (upper row), then rotated 180° (lower row). The age distribution of the 5150 otoliths in the training (blue), and the 515 otoliths in the test (orange) set (bottom).

2.2 Convolutional neural network architecture

145

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

155

CNNs were selected based on performance on the ImageNet benchmark and the availability of open-source implementations with pre-trained weights. The CNN models are aimed at classification, while we treated aging as a regression problem (Moen et al., 2018; Vabø et al., 2021). The last layer of the CNNs was therefore modified to a linear output. For the EfficientNetV2 family we did this by applying three multi-layer perceptron layers going from 1280 output of the last hidden layer to a dense 256-layer, then a leakyRelu (Xu et al., 2015) layer, then a dense 32-layer, then a leakyRelu layer, and finally a linear output layer. For EfficientNet we only changed the last layer from softmax to a linear output (Figure 8 in supplementary materials).

164

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

167

2.3 Implementation and training

168

EfficientNetV1 B4, B5, and B6 were imported and modified with TensorFlow (Abadi et al., 2016) and Keras (Chollet and others, 2018) software packages in Python. Computation was done using CUDA 11.1 and CuDNN with Nvidia (Nvidia Corp., Santa Clara, California) A6000 accelerator card with 48 GB of GPU memory and P100 cards with 12 GB of GPU memory, EfficientNetV2 Medium, and Large were imported and modified with the PyTorch (Paszke et al., 2019) and Timm (Wightman, 2019)

174

software packages. Computation was done on P100 and RTX 3090 with 24 GB of GPU
175 memory. Pretrained weights for EfficientNet were available from Keras, and pre-trained
176 weights for EfficientNetV2 were available from Timm. The models will be referred to as
177 B4-Min, B4-Middle, B4-Max, B5-Min, Medium-min and so on by combining model
178 name with image exposure.
179

Augmentation is a commonly used technique to artificially inflate the training data
180 set by applying transforms that modify the input while preserving class. The images
181 were augmented using rotation between 0 and 360 degrees, and reflection by the vertical
182 axis.
183

The cost function used was mean squared error (MSE) while the metric used for
184 evaluating the models and comparing them to expert readers was accuracy. Accuracy
185 was obtained by rounding the real valued predictions to the nearest integer and
186 measuring the fraction of otoliths where the age classification matches the labels. It
187 should be noted that this measure is different from that commonly used in otolith
188 studies, where accuracy relates to the closeness of the age estimate to the true value
189 validated age, e.g. with radiocarbon methods Campana (2001).
190

The data set of 5150 otoliths was divided into a training set constituting 90% of the
191 otolith images (4635 otoliths) and a test set of 10% (515 otoliths). To get the most out
192 of a small data set we applied 10-fold cross-validation on the training set. The data set
193 is divided into ten parts, and in each iteration (or "fold"), a different part is retained for
194 validation, while the model is trained on the remaining nine parts. In other words, 10
195 different models were trained with a different set of 463 images used for validation in
196 each fold, i.e. each data point participates in the validation set once and in the training
197 set 9 times. Among the 10-fold models, the one with the best MSE was chosen. The
198 best model parameters on the validation set were then used to predict the age on the
199 test set, and the metric for accuracy and MSE were recorded. The test set is chosen at
200 random, while the 10-fold split of the training set is chosen using a stratified k-fold split,
201 which preserves a similar distribution of the whole cross-validation set in each validation
202 set. That means the 463 images in the validation set will have similar age distribution
203 to that of the 4635 images in the cross-validation set.
204

2.4 Hyperparameters

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

EfficientNet used learning-rate with a weight decay scheduler, while EfficientNetV2 used Cosine Annealing scheduler (Loshchilov and Hutter, 2016). The training- and validation image size used was not changed, except for EfficientNetV2 Large which uses a smaller validation image size. The exact configuration of each network is available with each network result on the GitHub page of the project (<https://github.com/emoen/Deep-learning-for-regression-of-cod-otoliths>). The hyperparameters are available in Table 3, and 4 in supplementary information.

2.5 Ensemble learning with averaging

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

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

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

2.6 Correlations of predictions on the test set and clustering analysis

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

2.7 Comparison of CNN with human readers

To evaluate the credibility of CNN predictions in relation to human readers, we compared the mean percentage agreement of the test set predictions within each age class with those from multiple human readers from a recent internal cod age reading workshop carried out in 2021 at the Institute of Marine Research, Norway. In this workshop, a set of 100 broken otoliths from Atlantic cod were read by seven readers, of which five were certified advanced cod readers and two were under training. By comparing the results of the test set to the mean agreement and standard deviation of predictions within the age class from the workshop, we evaluated if machine-driven estimates were behaving in line with those anticipated by human readers.

3 Results

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

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

The mean MSE of the 17 models was 0.284 on the test set, and the standard deviation was 0.022. The highest MSE was from B5-max with MSE of 0.359, and the lowest MSE was from B6-middle exposure with MSE of 0.262. The models were

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

statistically different (ANOVA, $p = 1.6 * 10^{-7}$), but these differences were not
263 significant for the individual factors of model architecture (two-way ANOVA, $p = 0.139$)
264 or image exposure ($p = 0.057$). See Table 13 in the supplementary information for a
265 T-test of all models. From the interaction plot for two-way ANOVA, we see that using
266 low exposure images is more beneficial for the smaller models (Figure 2).
267

Medium and Large were the best models with a MSE of 0.278, and the 9-channel
268 composite images gave better results than any individual exposure, with a MSE of 0.272.
269 The high MSE for B5-max and B6-max was due to a large misprediction of the image
270 with index 308 in the test set labeled 1 year and predicted 5.7 years (see Table 7 in
271 supplementary information on outliers).
272

Medium-all was the highest scoring model with percentage agreement (PA) 91.3%
273 and B4-max was the lowest scoring model with PA 87.6% (Table 1). Medium was the
274 overall best performing model and B4 was the worst. The 9-channel composite images
275 outperformed individual exposures, while max exposure had inferior performance to the
276 others.
277

When comparing each 10-fold ensemble average prediction accuracy, and MSE for all
278 17 models, the ensemble metric was either better than or in the upper quantile for all
279 the models (Figure 3). The prediction MSE and accuracy of each fold are given in
280

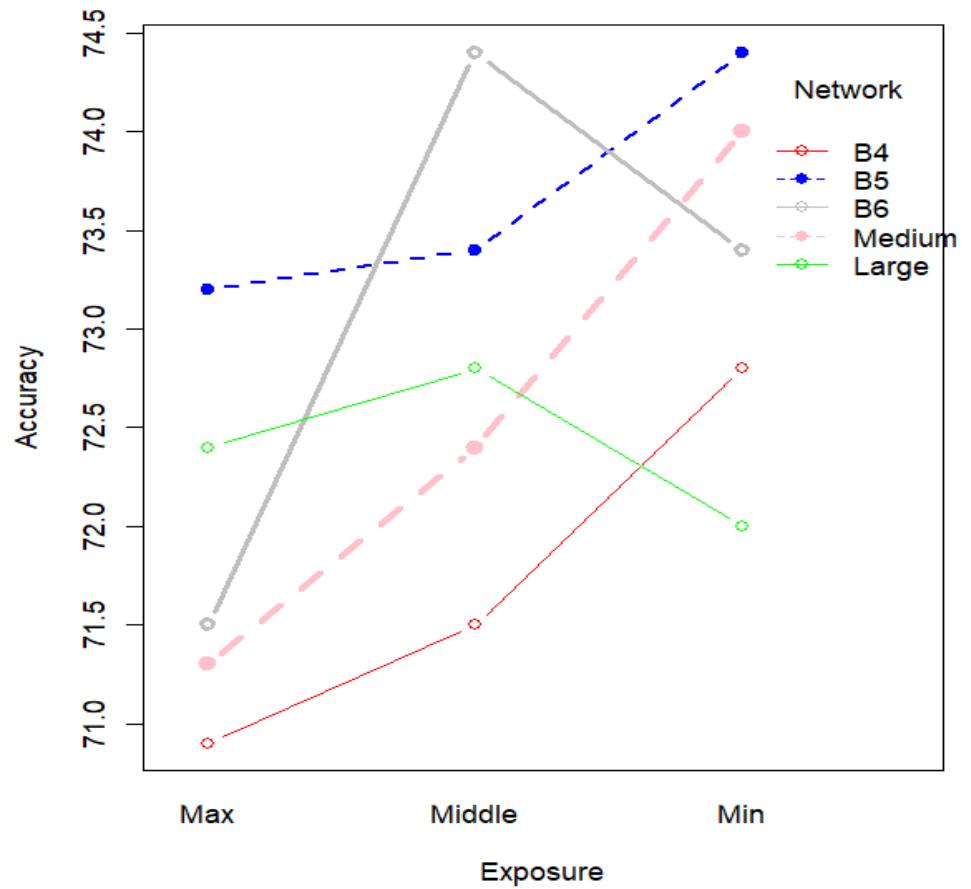


Figure 2. Interaction plot of the 5 networks with image exposure on x-axis and ensemble accuracy on the y-axis. We see that under-exposed images perform better for all but the B6 and Large network.

supplementary information (Table 5, 6).

281

3.1 Prediction by age class

282

When calculating the accuracy of all models by age class, we found that accuracy for
283 one- and two-year-old was the highest at more than 90% (supplementary information in
284 Figure 10). All otoliths aged 1 to 6 were correctly classified with more than 70%
285 accuracy, while older fish had varying degrees of accuracy. The few 13-year-old were for
286 example all predicted to be younger.
287

288

No systematic bias in the age prediction of CNN is visible except for the
288 underestimated age of individuals aged by the expert reader as 13 years old (Figure 4).
289

290

3.2 Simple ensemble-average predictions

290

We searched the space of ensembles-average predictions of 2 to 17 models, which is the
291 set of unordered combinations without replacement, equal to the binomial coefficient
292 $\sum_{k=1}^N \binom{N}{k}$ where $N = 17$ and $k \in 2.N$. For each set of ensemble combinations, we
293 recorded the best ensemble and found that the best overall ensemble prediction was an
294 ensemble of six models which produced an accuracy of 78.6%. The ensemble consisted
295 of B4-min, B5-min, B6-min, Medium-min, B6-middle, and B4-max. The results are
296 presented in detail in supplementary information in Table 10, 11, and 12.
297

298

The ensemble accuracy decreased after adding 6 models while the MSE continued to
298 decrease until all 17 models were included, which was as expected from the theory on
299 simple ensemble average learning since the variance is reduced with more models.
300

301

The models B4-min (No 1) and B6-min (No 3) were those most often present in the
302 top scoring model with inclusion in 14 ensembles (Table 2). These models did not have
303 the highest accuracy (B5-min, and B6-middle) but an accuracy of 72.8% and 73.4%.
304 This was lower than the highest accuracy models, which were B5-min and B6-middle
305 (74.6%) with a rank of 3 and 5, respectively.
306

307

The mean ranking by exposure types was: min-exposure (rank 4.4), middle-exposure
307 (rank 8.6), 9-channel composite (rank 10), and max-exposure (rank 11.2). The mean
308 ranking by architecture was EfficientNet (rank 6.6), and EfficientNetV2 (rank 10.3).
309

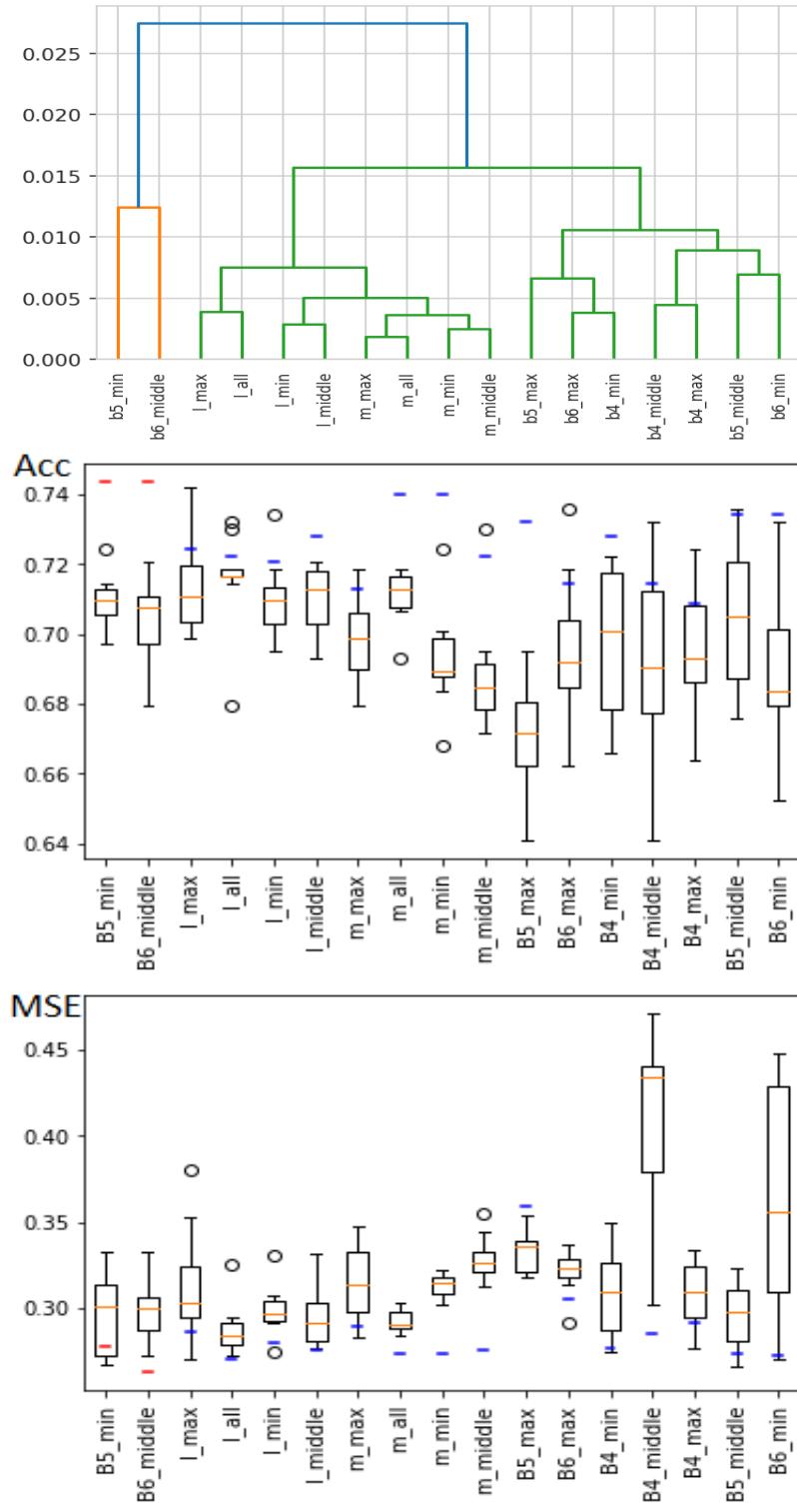


Figure 3. Hierarchical clustering (HCA) on the correlation of predictions (top), a box-plot of accuracy score (middle), and MSE (bottom) of all the 17 models. In (middle) and (bottom), the blue line is ensemble-average prediction accuracy (or MSE) on the test set, the red lines are the two best ensemble-average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

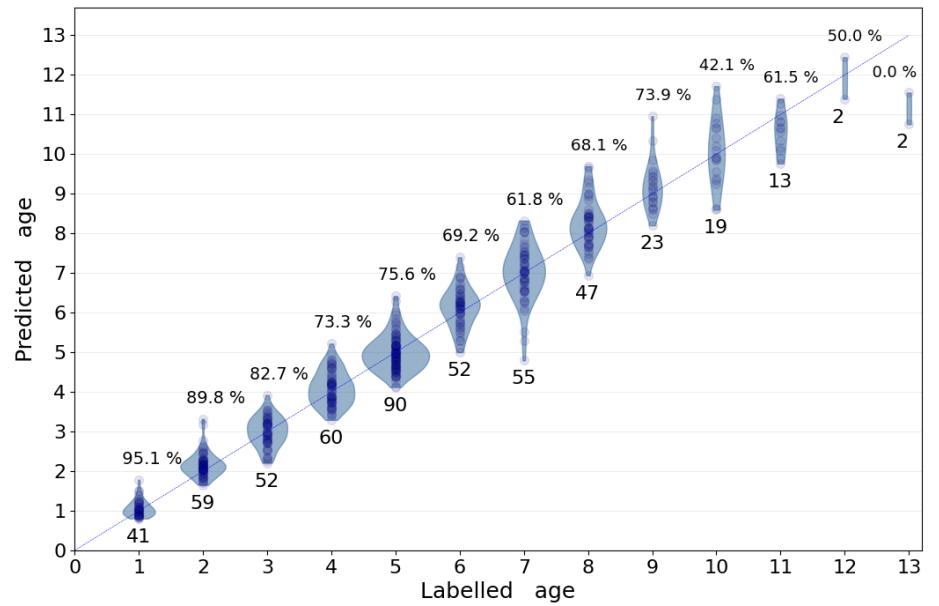


Figure 4. Violin plot of predicted age from model B5-min with accuracy of 74.4%. Above each age is the accuracy, and below is the total number of images in the test set of that age class

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

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 changed drastically or when break lines interrupted the normal pattern of ring deposition. In one case (image 362), all models estimated the otolith to be 5 instead of 7 years old: upon visual investigation, the otolith was clearly 5 years old, and the initial age had likely been misread.

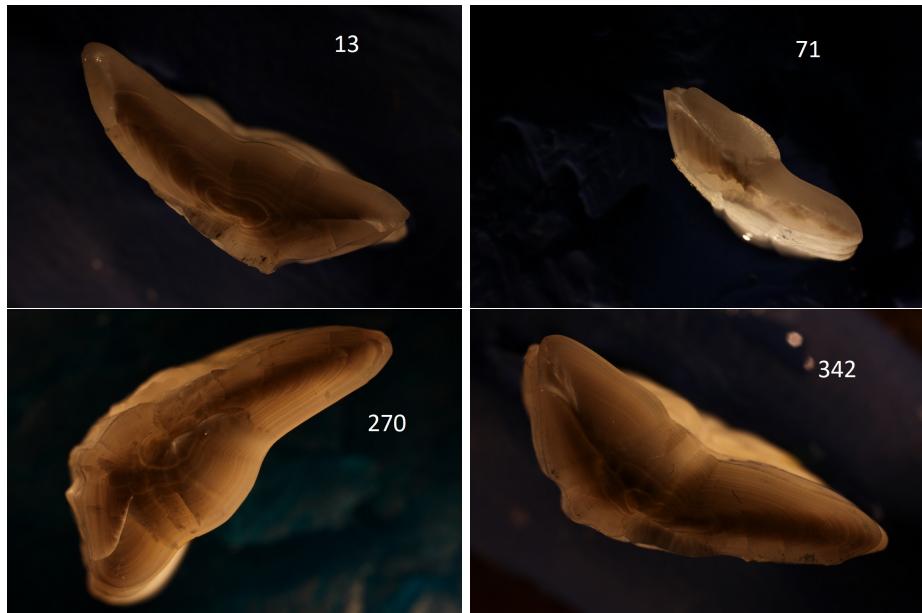


Figure 5. Example outlier images with index 13, 71, 279, 342 from the test-set were mispredicted by between 25% and 100% of the models

Some cod otoliths were outliers to all models and on all exposures (e.g. otoliths 71,

342, 362, and 369), to a family of models and on all exposures (e.g. otoliths: 13, 423),
328
to some models and on one exposure (E.g otolith 308), and to both families of models
329
and on some exposures (*E.g.* otolith 320).
330

We also observed that the number of outliers did not correlate with model
331
performance. *E.g.*, B5-min, and B6-mid which had 7 and 9 outliers, but the best
332
accuracy. While B4-max with the lowest accuracy (70.9%) had the least number of
333
outliers with only 6 mispredictions.
334

3.4 Correlation of predictions and cluster analysis

335

The correlation of models on the test-set predictions given in Figure 11 in
336
supplementary information shows that the models strongly correlate in outlier
337
predictions. The correlation from all the predictions on the test set varied between
338
0.988 to 0.999, with the lowest correlation found between B5-min and Medium-min.
339

Hierarchical clustering (HCA) of the models found 3 clusters. One cluster contained
340
B5-min and B6-middle, which were the two best performing models. A second cluster
341
contained of all the EfficientNetV2 models, and a third cluster contained the rest of the
342
models (Figure 3, and 11).
343

The two least correlated models, B5-min and Medium-min, which had Pearson's
344
correlation of 0.988 showed strongly correlated predictions also on a sub-year scale
345
(Figure 6).
346

3.5 Comparison of CNN with human readers

347

Variations in percentage agreement between age classes showed similar patterns in both
348
CNN-based predictions and human readers, with generally decreasing agreement with
349
age (Figure 7). Within each age class, percentage agreement from CNN-based
350
predictions was lower than the average for multiple human readers and increasingly so
351
for the older age classes. However, they often remained within or close to the range of
352
percentage agreement observed across all readers for all otoliths of a given age class.
353

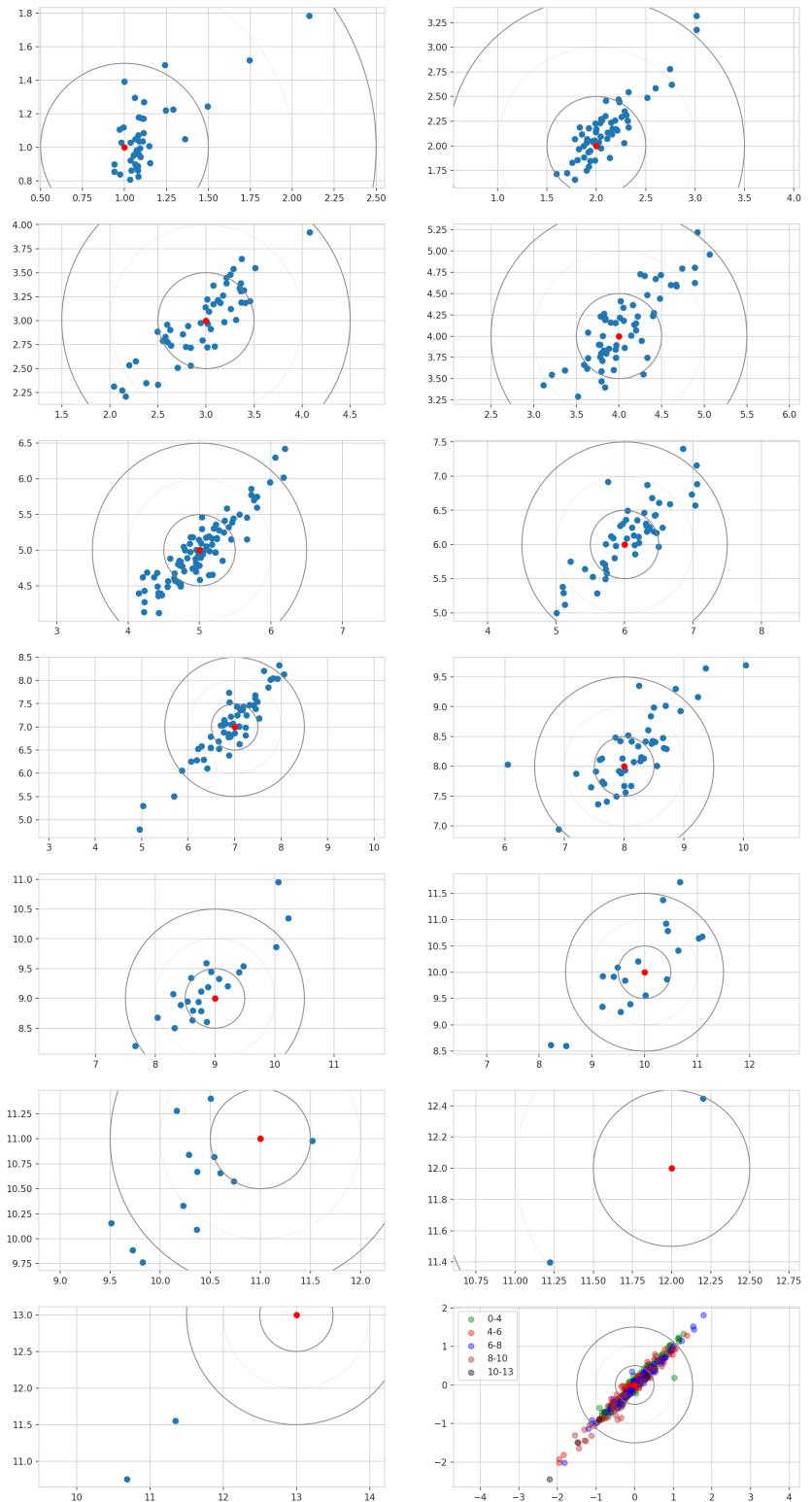


Figure 6. Comparison of age estimates predicted by Medium-min (x axis, years) and B5-min (y axis, years) as age-specific scatter plots, and in aggregate for all age groups in the bottom right panel. The circles show age differences of 0.5 and 1.5 years.

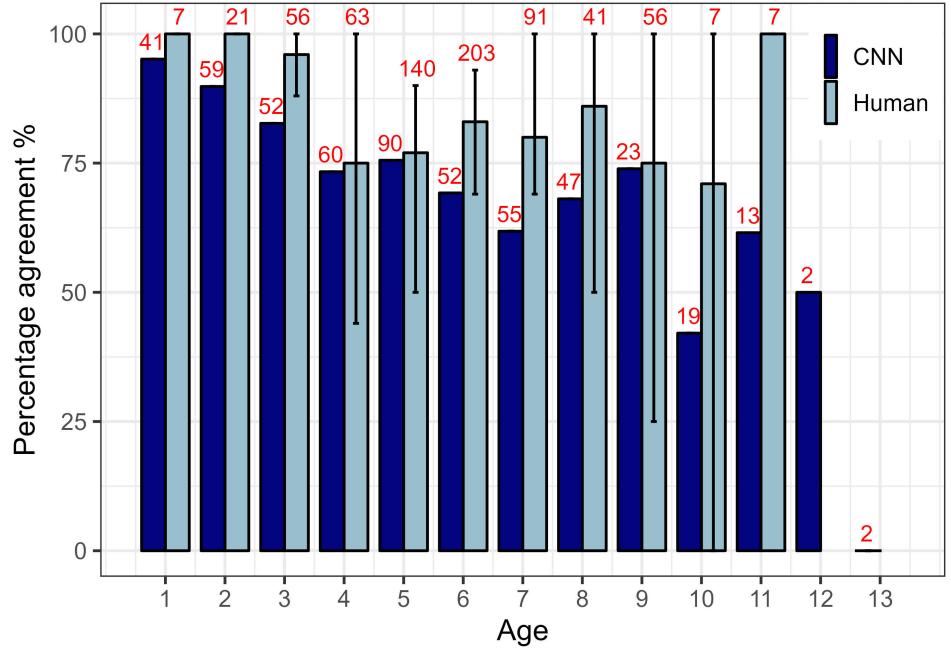


Figure 7. Comparison of mean percentage agreement within each age class for two sets of otoliths: the CNN-predictions on the test set (black); an internal age reading of 100 cod otoliths involving 7 readers (gray). Numbers indicate the total number of readings for each age class (with 1 reading per otolith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

We successfully trained ~~and tested machine learning methods~~ convolutional neural networks (CNNs) on images of broken ~~otoliths~~, and achieved a maximum accuracy of 78.6% with an ensemble consisting of six models, noting that the accuracy is the fish otoliths. The CNNs were tested against fish age classifications made by human readers and achieved an accuracy (agreement between the read ages and the model predictions) of 78.6% using an ensemble consisting of six models.

4.1 Accuracy across different age groups

The age of the younger individuals was predicted with greater accuracy ~~using the CNN models~~ than those of older individuals ~~by our models~~. Thus, the CNN appears to be particularly competent at aging cod otoliths of younger age ~~classes~~. 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
368 same. Human expert readers use various visual cues, prior knowledge, and background
369 information to determine fish age, such as comparing ring counts on multiple axes and
370 having intrinsic knowledge of the periodicity of opaque and translucent zones for a given
371 species. In younger fish, the increments are usually wider and more clearly separated as
372 fish -and consequently otolith- growth rates are maximal prior to maturity. Fish of age 1
373 are small and have comparatively small otoliths with a straightforward ring pattern
374 made of one single finished opaque and translucent zone, and expert readers are unlikely
375 to disagree on its interpretation. On the other hand, a CNN architecture as used here
376 identifies hierarchical patterns on different scales of the image from which it derives a
377 value in the range of those provided in the training set. This means that unless
378 specifically forced to do so, the algorithm may seek and interpret visual clues other than
379 the rings human readers are trained to use. A possible explanation for the higher
380 prediction accuracy of younger fish is that age is related to the area the otolith covers
381 relative to the total image size. Because the same camera settings were used, all images
382 had the same dimension and calibration. For a species with moderately large adults
383 such as Atlantic cod (Froese and Pauly, 2022), the otoliths will grow in size significantly
384 faster during the first years and then slow down with approaching sexual maturity. As
385 fish get older different growth trajectories will then lead to greater overlap in otolith
386 sizes across different ages. It is therefore possible that the CNNs are not counting
387 growth zones as human expert readers would, but rather that they synthesize all
388 available patterns in the image to find recurring ~~characteristics to the ages provided~~
389 age-related characteristics evident in the training set. The size of the area that the
390 otoliths cover against the more uniform black background might for example be a very
391 simple feature picked up by the CNNs ~~yet~~ with high predictive power for the youngest
392 fish. , while the higher inter-individual variability and greater size overlap at older ages
393 would affect the predictive accuracy of CNNs
394

The hypothesis that CNNs exploit other information than the growth zones is
395 consistent with the findings of an earlier study where network activations inside a CNN
396 were explored for images of Greenland halibut otoliths (Ordonez et al., 2020).
397 Visualisation techniques were used to reveal the relative importance of attributes such
398 as shape, inner structure, and size of the otoliths using activation maps. Importantly,
399

the authors found that the CNN utilized information in pixels corresponding to annual
399 increments to only a small extent. To explore this possibility, we attempted to train a
400 network using otolith silhouettes only, e.g. images where all internal structures were
401 erased. We could not achieve acceptable performance of the CNN models ~~in~~during
402 these initial tests.
403

4.2 Sub-year agreement between models

One surprising observation was that models agreed with each other on a sub-year scale
405 to a remarkable degree. One might expect that model output is drawn from a Gaussian
406 or symmetric distribution around the correct (integral) value, possibly with some bias.
407 Instead, we saw that the different models classify individual otoliths with high sub-year
408 agreement. As the labels are integral values, the models must infer this fractional age
409 from some characteristics of the input.
410

4.3 Importance of training set size relative to model 411 performance

It is commonly recognized that the performance of deep learning systems often improves
413 with more training data (LeCun et al., 2015). A crucial issue in machine learning
414 projects is then determining the amount of training data needed to achieve a specific
415 performance goal. In this study we utilized a somewhat large data set of around 5000
416 images, although the images were divided among a large range of age classes. In
417 comparison, it is not uncommon for deep learning systems used for image classification
418 such as ImageNet to be trained on thousands of images for each class (Russakovsky
419 et al., 2014). In this study, the use of transfer learning (Yosinski et al., 2014) and
420 augmentation yielded a significant performance boost but it is still likely that the
421 network would provide more robust predictions with a larger training set. ~~From-a~~
422 ~~preliminary initial training~~ During preliminary analysis, not reported as part of this
423 study, we trained a B4 network on ~~about~~around 2000 images and obtained an accuracy
424 of around 60%. When another 3000 images were added to the data set accuracy reached
425 about 70%. This ~~could suggest that increasing our sample size would have further~~
426 suggests that further increases in sample size could have increased accuracy.
427

4.4 The effect of image size

428

The high-resolution 3744×5616 cod otolith images were scaled down to between 429
 380×380 and 528×528 pixels to match the requirements of the different EfficientNet 430
architectures. This reduction in resolution may have affected the readability of 431
finer-scale visual features such as growth rings. The fixed camera setup resulted in the 432
background constituting a large proportion of the images, especially for smaller otoliths 433
from the youngest individuals. This is especially true due to the curved or oval shape of 434
the otolith, as a compressed image will not only have less pixels to work with but will 435
also have a comparatively more important fraction of black background, which is 436
effectively useless for age interpretation. Otoliths of other fish species like red mullet 437
which have a more circular shape may be less sensitive to this problem (Politikos et al., 438
2021). Improvements might therefore be made by instead first isolating the otoliths 439
from their background and cropping the image accordingly, in order to have a machine 440
learning network trained on using exclusively before training the network on the 441
information contained exclusively within the area of interest. This would also limit 442
information loss from image compression. 443

4.5 Outliers and transparency

444

The different networks were generally able to predict the age of otoliths with CNNs 445
trained in this study produced predictions of fish age that were generally within less 446
than a one-year deviation from year off the labelled values. It is noteworthy that 447
predicted ages were similar across different models and errors of more than a year were 448
only seen in 2% of the predictions on made using the test set. Closer inspection revealed 449
that such errors were often caused by otolith images with poor readability, in particular 450
drastic changes in exposure or visual damages and interruptions on the reading axis. 451

Interestingly, for one of these images, the predicted age was correct, and a 452
reexamination re-examination by an expert revealed that the initial annotation was 453
wrong by two years. While the previous results suggest that the network may not have 454
relied entirely on ring patterns for estimating age, this correct prediction of a wrongly 455
assigned age shows that it is still utilizing cues that are somewhat age-specific. Model 456
behaviour was also similar across all networks on single predictions of outliers: four were 457

identified in all of the models, suggesting they must have learned the same features.

458

4.6 Effects of image exposure on predictive power

459

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

468

4.7 Effects of 9-channel composite images and architecture size

469

~~Combining all 3 exposures into a 9-channel image did not perform better even if We found that CNN model performance did not improve when more information was available to these CNNs made available to the network by combining all 3 exposures into a 9-channel image.~~ The EfficientNetV2 models trained on these images performed similarly to models trained on single-exposure images. The variance of predictions on 9-channel images was noticeably lower than for regular images, meaning the CNNs were more certain in their prediction even when the prediction was wrong.

476

We also found that the newer and larger EfficientNetV2 architecture did not stand out as better than the EfficientNetV1 models. On the contrary, some of the best models were the smaller ones of B4, B5, and B6. This could be due to ~~the size of our data set not being large enough to utilizing all the there not being enough information in our relatively small data set to fully utilise the large number of model~~ parameters in the larger models. Larger networks are generally able to better explore a larger data set, such as ImageNet, through training.

483

4.8 Utilising model ensembles

484

We observed slight improvements in performance when an ensemble of models was used
485 for prediction. The use of numerous models in ensembles resulted in large numbers of
486 combinations of model predictions with varying accuracy. Some combinations achieved
487 higher accuracy than others (close to 79% for some combinations of six and seven
488 models). However, the mean ensemble prediction accuracy for a given number of models
489 showed that five models or more in combination resulted in accuracy just above 75%.
490 Five models thus seem to be sufficient and there could be minimal gain in precision in
491 combining larger numbers of models. Interestingly, ensembles combining models with
492 higher variance resulted in better predictions. This ~~can~~may indicate that if models are
493 too similar in individual predictions, the averaging effect ("wisdom of the crowds") will
494 not play out in the same way as when models with higher variance are combined.
495 Remarkably, many predictions only disagreed with a small decimal fraction. This could
496 imply that the models learned the same features in the otolith images.
497

4.9 Comparison of CNN with human readers

498

The comparison of age-specific percentage agreement in CNN-based predictions with
499 those from an internal age reading workshop showed that our models may achieve
500 similar agreement with human readers. While the numerical results are not directly
501 comparable in the sense that two different sets of otoliths were read, the trends in mean
502 precision across age groups were similar. Of particular interest is the fact that the mean
503 percentage agreement for our CNN-based predictions for a given age group generally fell
504 within or close to the range of percentage agreement for all otoliths of a given age class
505 seen between all readers involved in an age reading workshop. This may indicate that
506 while machine-based methods may not yet have the predictive accuracy of an expert
507 human reader, their estimates still fall within the expected range and may not be easily
508 distinguishable from those of traditional readers. Further testing should be conducted
509 to assess whether this is consistent, for example, by conducting a multi-reader aging
510 event that includes undisclosed machine-based estimates of the same samples and
511 monitoring how they compare and whether they can be picked out by human readers.
512

4.10 Resource efficiency

513

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

533

An additional advantage of the setup developed here is its high transferability. For the majority of ML-algorithms, standardization of training material is essential to ensure it can be transferred across. By using a simple setup comprising a mounted DSLR with a macro lens and an external light source, instead of a more costly and specific microscope camera, ~~ours is easily transferable~~ facilitates repeatability: any camera fitted with a similar focal length and using the same acquisition parameters would give identical images.

540

~~Also, CNNs can~~ CNNs can also be applied without high additional cost or even be incorporated into ~~the routine protocols, but add a new~~ routine protocols and provide additional value e.g., reading consistency check, time-drift evaluations, inter-reader

543

comparisons (how much each reader is ‘off’ ~~is each reader~~ when compared to ~~the~~ CNN 544 predictions, even if not compared with the same otolith samples), etc. ~~The advantages 545 of ensemble predictions will also be easier to gain with networks. Ensembles of several 546 expert readers are highly resource demanding—especially when scaling to huge 547 datasets—while an ensemble of, say eight versus three networks only requires a little bit 548 more computation.~~ Also, networks are comparatively easier to scale up than the 549 number of human expert readers, especially when analysing huge datasets, as any 550 increase in ensemble size (number of networks) only requires an increase in computer 551 resources. 552

We see the process of CNN implementation as an evolution of the protocols, with 553 an intensive phase of model development and training. Through gradual improvement 554 of model reliability, CNNs could emerge as a complementary supportive tool for 555 traditional age estimations. The integration of those technologies could help scale the 556 capacity of age reading experts and improve the sampling of biological data and 557 monitoring of various fish stocks. 558

While the exact features used by the networks may differ from those interpreted by 559 human readers, one may be content with a trained network as a black box relying 560 entirely on its empirical accuracy. Deep learning techniques are particularly powerful in 561 detecting patterns in data (LeCun et al., 2015), and whether the networks actually 562 detect and count annual increments as the defining features or not, a causal relationship 563 between what the network attends to and the structure of the otoliths is likely. If the 564 network does not use growth zones as the primary features for prediction, this means 565 that it has found useful correlating patterns unavailable or not obvious to the human 566 eye. This was demonstrated in the case of Greenland halibut otoliths where the shape 567 of the otoliths seemed to be the defining characteristic correlated with age (Ordonez 568 et al., 2020). Machine learning frameworks may therefore be used as complementary age 569 readers to for experts participating in otolith image interpretation workshops, as 570 informative input playing part of additional background information, or relied upon as 571 autonomous age readers without a subjective bias specific to increment identification 572 and counting. 573

We see the process of CNN implementation as an evolution of the protocols, with 574 an intensive phase of model development and training. Through gradual improvement 575

of model reliability, CNNs could emerge as a complementary supportive tool for
576 traditional age estimations. The integration of those technologies could help scale the
577 capacity of age reading experts and improve the sampling of biological data and
578 monitoring of various fish stocks.
579

4.11 Conclusion

Our results demonstrate that deep learning techniques have ~~a major potential for~~
581 ~~otolith analysis~~huge potential in extracting age information from otolith images.
582

Standard model architectures trained on sufficient training data specific to the use case
583 can accurately predict age from images of broken otoliths. We believe that carefully
584 trained CNNs could become a major component in procedures that require minimal
585 processing and could be able to produce near at sea age estimates. In addition,
586 algorithmic age estimates could serve as a useful reference for evaluating age reader
587 biases within and between otolith laboratories.
588

When developing the CNN framework for the automatic age estimation, we found
589 that the B4 ~~architectures were~~ model was quick to train and ~~that they~~ performed well.
590 Ensemble approaches can also be considered if the increased computing effort is not a
591 constraint, as they can provide more robust and higher-performing predictions. For a
592 quick-to-train ensemble, ~~the~~ B5 and Medium models might be added. Our results also
593 indicate that the use of slightly under-exposed images may ~~be beneficial~~improve model
594 performance.
595

5 Acknowledgements

We thank Jane Godiksen and age readers and technicians from the Demersal Fish
597 research group for providing otolith age estimates and images used for this study. We
598 thank Erlend Langhelle for providing insight into the image-taking-protocol and age
599 interpretation of cod otoliths. We thank the anonymous reviewers for their contribution
600 to improving and clarifying the manuscript.
601

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.*
- 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.
Fisheries Research, 220: 105351.
- 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
developments in fish otolith research. *University of South Carolina Press, Columbia,*
S.C., pp. 545–565.
- 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.
- 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
biology*, 59(2):197–242.
- Carbonara, P. and Follesa, M. C. (2019). Handbook on fish age determination: a
mediterranean experience. *General Fisheries Commission for the Mediterranean.
Studies and Reviews*, 98: 1–179.
- Chollet, F. and others (2018). Keras 2.1.3. <https://github.com/fchollet/keras>.
- 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
and exploitation. *Global Change Biology*, 26(10):5661–5678.
- Fisher, M. and Hunter, E. (2018). Digital imaging techniques in otolith data capture,
analysis and interpretation. *Marine Ecology Progress Series*, 598:213–231.
- 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*.

NRC Research Press Ottawa, Canada.	629
https://cdnsciencepub.com/doi/abs/10.1139/f04-063 (Accessed 3 February 2022).	630
Froese, R. and Pauly, D. (2022). Fishbase.	631
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.	632 633 634 635
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.	636 637 638
ICES (2013). Report of the second workshop of National Age Readings Coordinators (WKNARC2). In <i>Report of the Second Workshop of National Age Readings Coordinators (WKNARC2)</i> . ICES, Copenhagen.	639 640 641
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 25</i> , pages 1097–1105. Curran Associates, Inc.	642 643 644 645
LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. <i>Nature</i> , 521(7553):436–444.	646 647
Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. <i>neurips</i> .	648 649
Mahé, K. (2009). Project no. 044132. <i>Automated FISH Ageing (AFISA): final activity report</i> .	650 651
Mingxing Tan and, Q. V. L. (2021). Efficientnetv2: Smaller models and faster training. <i>CoRR</i> , abs/2104.00298.	652 653
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> .	654 655

- Myers, S., Thorsen, A., Godiksen, J., Malde, K., and Handegard, N. (2019). An efficient 656 protocol and data set for automated otolith image analysis. *GeoScience Data Journal*. 657
- Ordonez, A., Eikvil, L., Salberg, A.-B., Harbitz, A., Murray, S. M., and Kampffmeyer, 658 M. C. (2020). Explaining decisions of deep neural networks used for fish age 659 prediction. *PLoS ONE*, 15(6):e0235013. 660
- Panfili, J., de Pontual, H., Troadec, H., and Wright, P. J. (2002). Manual of fish 661 sclerochronology. <https://archimer.ifremer.fr/doc/00017/12801/> (Accessed 3 662 February 2022). 663
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, 664 Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, 665 M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. 666 (2019). Pytorch: An imperative style, high-performance deep learning library. In 667 Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, 668 R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 669 Curran Associates, Inc. 670
- Politikos, D. V., Petasis, G., Chatzispyrou, A., Mytilineou, C., and Anastasopoulou, A. 671 (2021). Automating fish age estimation combining otolith images and deep learning: 672 The role of multitask learning. *Fisheries Research*, 242:106033. 673
- Ragonese, S. (2018). Methuselah or butterfly? when fish age estimates and validations 674 tell different stories. the case of the european hake (*merluccius merluccius* l. 1758) in 675 the mediterranean sea. ””. 676
- Reglero, P. and Mosegaard, H. (2006). Onset of maturity and cohort composition at 677 spawning of baltic sprat *sprattus sprattus* on the basis of otolith macrostructure 678 analysis. *Journal of Fish Biology*, 68: 1091–1106. 679
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., 680 Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). 681 Imagenet large scale visual recognition challenge. 682
- Sigurdardóttir, A. R., Sverrisson, D., Jónsdóttir, A., Gudjónsdóttir, M., Elvarsson, 683

- B. D., and Einarsson, H. (2023). Otolith age determination with a simple computer vision based few-shot learning method. *Ecological Informatics*, 76:102046. 684
- Siskey, M. R., Wilberg, M. J., Allman, R. J., Barnett, B. K., and Secor, D. H. (2016). 685
Forty years of fishing: changes in age structure and stock mixing in northwestern 686
atlantic bluefin tuna (*thunnus thynnus*) associated with size-selective and long-term 688
exploitation. *ICES Journal of Marine Science*, 73: 2518–2528. 689
- Smoliński, S., Deplanque-Lasserre, J., Hjörleifsson, E., Geffen, A. J., Godiksen, J. A., 690
and Campana, S. E. (2020). Century-long cod otolith biochronology reveals 691
individual growth plasticity in response to temperature. *Scientific reports*, 10(1):1–13. 692
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the 693
inception architecture for computer vision. *CoRR*, abs/1512.00567. 694
- Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional 695
neural networks. *CoRR*, abs/1905.11946. 696
- Tyler, A. V., Beamish, R. J., and McFarlane, G. A. (1989). Implications of age 697
determination errors to yield estimates. *ICES Journal of Marine Science*, 108: 27–35. 698
- Vabø, R., Moen, E., Smoliński, S., Åse Husebø, Handegard, N. O., and Malde, K. 699
(2021). Automatic interpretation of salmon scales using deep learning. *Ecol. Inform.* 700
63, 101322 (2021). 701
- Wightman, R. (2019). Pytorch image models. 702
- <https://github.com/rwightman/pytorch-image-models>. 703
- Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified 704
activations in convolutional network. *CoRR*, abs/1505.00853. 705
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features 706
in deep neural networks? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, 707
N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing* 708
Systems 27, pages 3320–3328. Curran Associates, Inc. 709

Supplementary information

710

Hyper-parameters

711

Table 3. Hyper-parameters on each model

Param/CNN	B4	B5	B6	Medium	Large
<code>train_batch_size</code>	8	8	16	8	8
<code>img_size</code>	380	456	528	384	384
<code>val_img_size</code>	380	456	528	384	384
<code>steps_per_epoch</code>	1600	1600	1600	1600	1600
<code>epochs</code>	150	150	250	450	450
<code>early_stopping</code>	-	-	-	40	40
<code>early_stopping_patience</code>	14	14	22	-	-
<code>reduceLROnPlateau_patience</code>	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
<code>learning_rate</code>	1e-05	v	v
<code>n_fold</code>	10	v	v
<code>test_size</code>	0.1	v	v
<code>in_chans</code>	3 or 9	v	v
<code>reduceLROnPlateau_factor</code>	0.2	v	x
<code>which_exposure</code>	min, medium, max	v	x
<code>scheduler</code>	CosineAnnealingLR	x	v
<code>T_max</code>	10	x	v
<code>min_lr</code>	1e-06	x	v
<code>weight_decay</code>	1e-06	x	v
<code>which_exposure</code>	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.

Model architecture

712

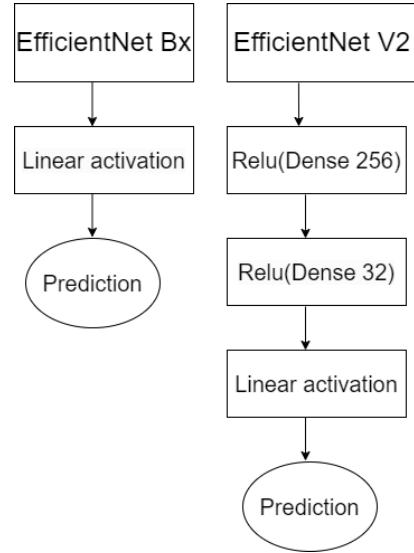


Figure 8. Diagram showing the changes made to the two architecture families

Descriptive Statistics

713

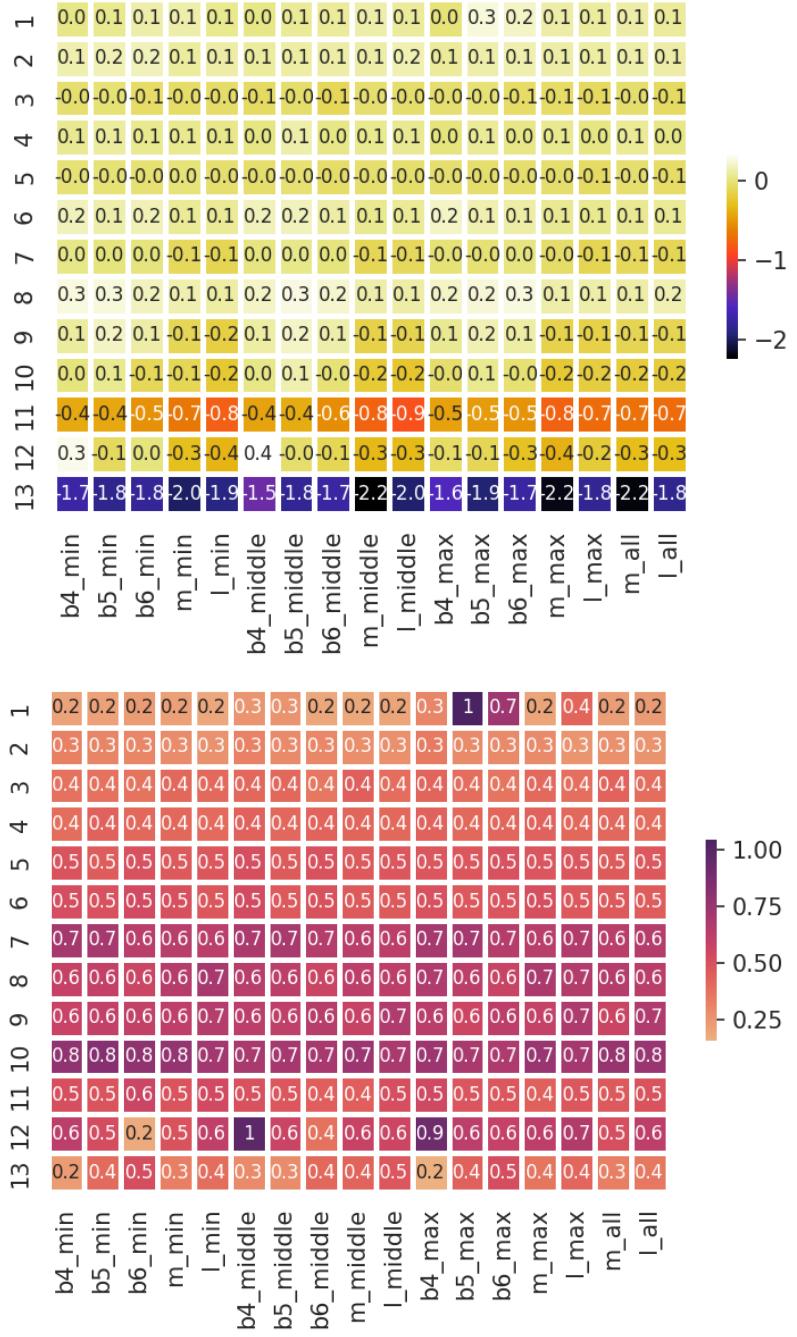


Figure 9. Model mean (top) and standard deviation (bottom) of residual test set prediction by age class

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

715

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

716

Predicted age class for all models and ground truth

717

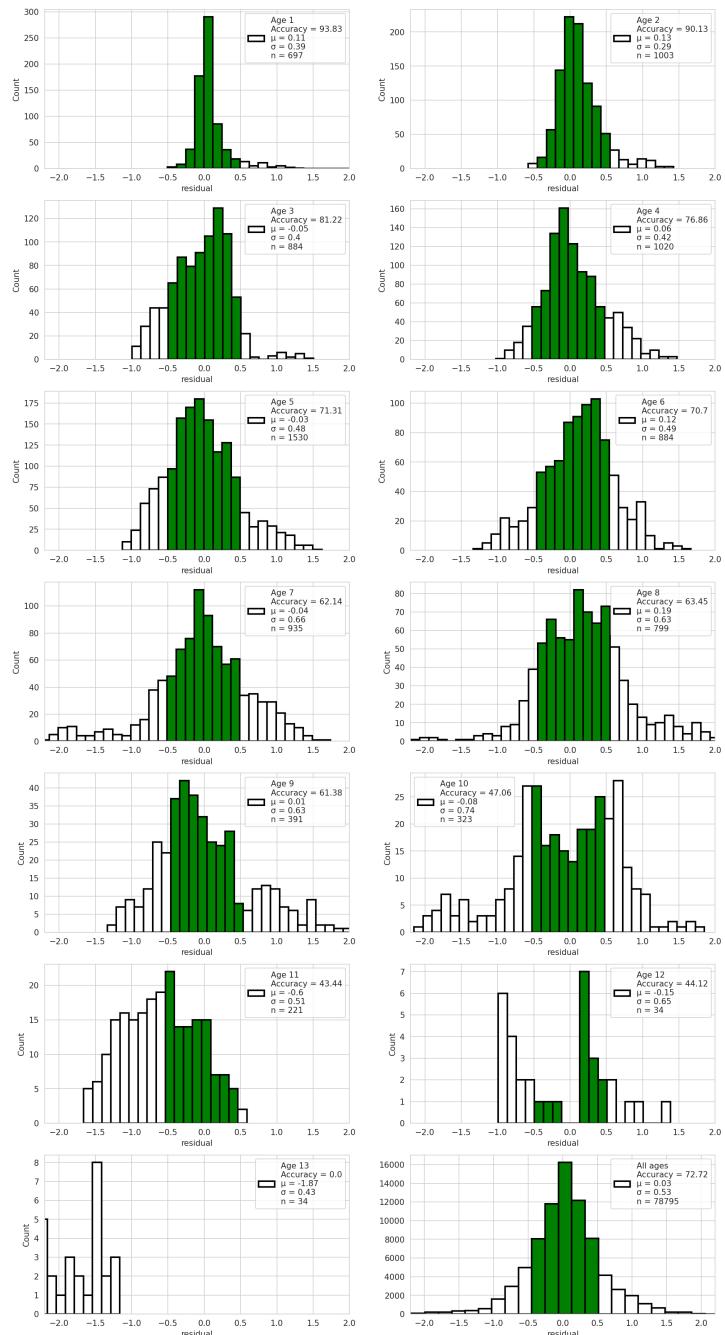


Figure 10. Predictions by age class from the average of all models. The green region shows the correctly classified age after rounding. The axis is fixed, hence outliers that differ from the true age by more than two years will not be visible.

Outliers

718

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						11.0					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					6.5					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				8.4			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		9.5			9.9	3.6	5.4			10.8	5.1	8.2					5.9				8.4			10
l-all		9.3			9.8					10.8	5.2	8.0	10.5				6.2				8.5			9
Age	8	8	6	7	13	7	10	8	1	11	7	6	13	7	10	9	11	8	11	5	10	11	-	
Count	9	2	1	1	17	7	1	4	16	3	2	2	17	17	6	2	7	2	1	3	3	3	141	
As pct	53	12	6	6	100	41	6	24	94	18	12	12	6	100	100	35	12	41	12	6	18	18	-	

719

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 were very bright on all three axes, and the dorsal axis had a break line (fissure or physical break), 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 were of good quality, but there 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 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. 720
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 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 were over-exposed. 721

Ensembles by simple average

722

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.

723

724

725

726

727

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)

728

729

730

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

Pearson correlation of each model on test-set predictions

731

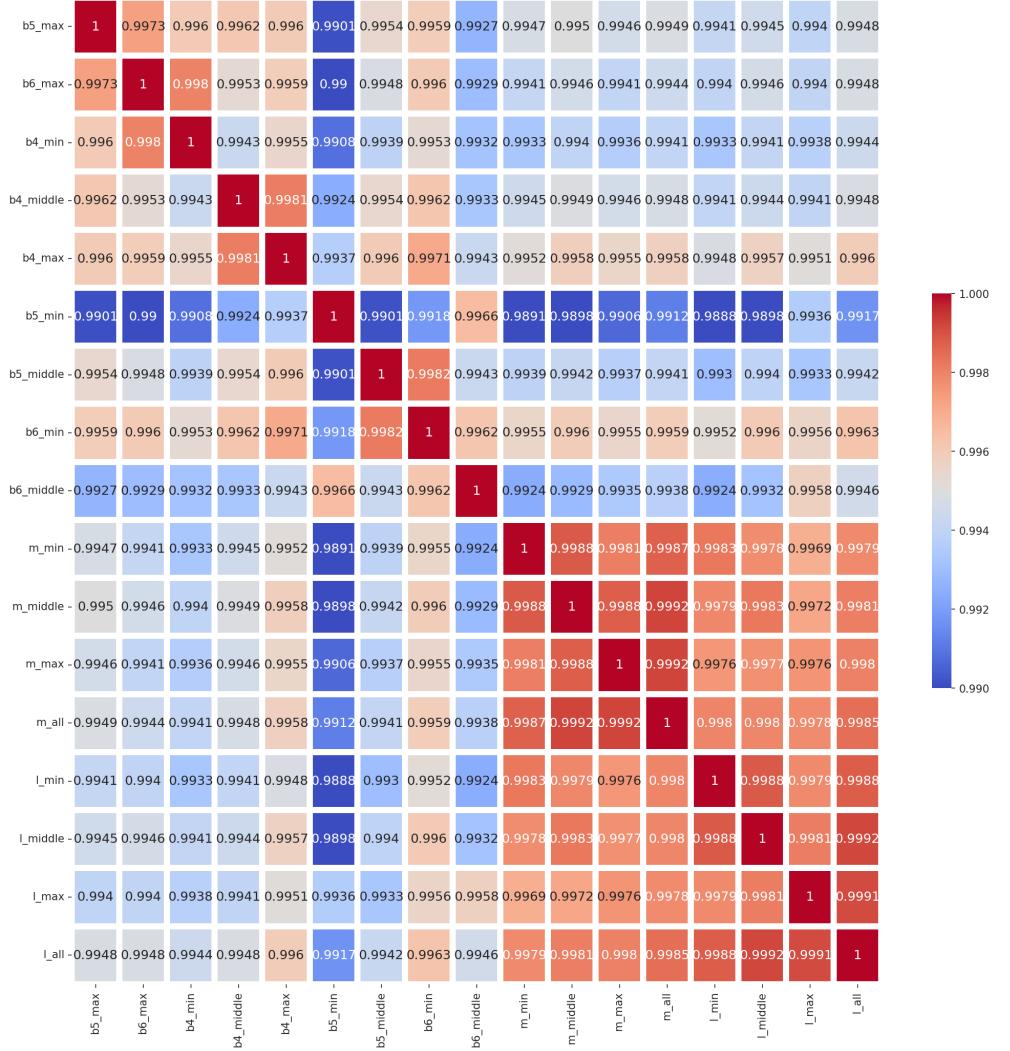


Figure 11. Pearson correlation of each model prediction on the test-set. The colors indicate the correlation coefficient, ranging from 0.99 (blue) to 1.00 (red).

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

732

prediction

733

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	0.0219	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	