

# Automatic interpretation of cod otoliths using deep learning

Endre Moen<sup>1\*</sup>, Rune Vabø<sup>1</sup>, Szymon Smoliński<sup>3</sup>, Come Denechaud<sup>1</sup>, Ketil Malde<sup>1,2</sup>,

**1** Institute of Marine Research, Bergen, Norway

**2** Department of Informatics, University of Bergen, Norway

**3** Department of Fisheries Resources, National Marine Fisheries Research Institute,  
Kołtajowa 1, 81-332 Gdynia, Poland \* endre.moen@hi.no

## Abstract

## Introduction

Knowledge of fish age structure is central to the study of fish and stock dynamics. 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 (Brunel and Piet, 2013; Hidalgo et al., 2011). Changes in the age distribution can track significant changes in population structure, such as a particularly strong year-class skewing the distribution (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, temporally resolved growth increments at the daily and annual levels (Albuquerque et al., 2019; Campana, 2001; Francis and Campana, 2011). 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).  
19  
Therefore, streamlining, scaling, and increasing the quality of age estimations can  
20  
improve the reliability of evaluations of fish biology and consequently stock size and  
21  
structure (Beamish and McFarlane, 1995; Ragonese, 2018; Tyler et al., 1989).  
22

Otolith reading is time and resource consuming. Training of expert readers can take  
23  
several years depending on the species, and otoliths often undergo a long processing  
24  
phase before the final age estimates can be produced (Carbonara and Follesa, 2019).  
25  
This is particularly true for demersal fish species, like Atlantic cod (*Gadus morhua*),  
26  
that have large opaque otoliths that typically require time-consuming preparation.  
27  
These routines vary between populations and institutes and range from direct reading of  
28  
broken otoliths under a magnifying glass, to embedding, thin sectioning and finally  
29  
imaging of the sections under a microscope. There has been a variety of methods  
30  
proposed to automatically interpret otoliths, which range from one-dimensional data  
31  
analysis like intensity transects (Mahé, 2009) to the more recent effort toward  
32  
developing machine learning (ML) frameworks (Moen et al., 2018; Politikos et al., 2021).  
33

## About deep learning and image analysis

34

Deep learning has during the last decade become the dominating field of machine  
35  
learning where various architectures of deep neural networks are able to learn to  
36  
efficiently identify patterns and structures in various types of data (LeCun et al.,  
37  
2015)(LeCun, 2015). Within computer vision, deep Convolutional Neural Networks  
38  
(CNN) have been prevailing the field ever since Krizhevsky et al. in 2012 (Krizhevsky  
39  
et al., 2012)(Krizhevsky et al., 2012) won the annual ImageNet Large Scale Visual  
40  
Recognition Challenge (ILSVRC) competition (Russakovsky et al., 2014)(Russakovsky  
41  
et al., 2015). CNNs have seen a continuous development with new improved  
42  
architectures arising year by year. ILSVRC remains the most important benchmark for  
43  
image classification, with 1.4 million images in the ImageNet training set. The  
44  
state-of-the-art CNNs are therefore heavily trained on a lot of images. Many of these  
45  
CNNs are publicly available including their trained network weights. This enables the  
46  
use of transfer learning from one image domain to another providing a very useful  
47  
pre-trained starting point for further training of more specific image classification tasks  
48

were much less images are available. Age estimation from images of otoliths represents, 49  
for several fish species, precisely such a task. InceptionV3 (Szegedy et al., 2015) was 50  
modified to predict the age of Greenland halibut (*Reinhardtius hippoglossoides*) from 51  
otolith images (Moen et al., 2018), and a modified InceptionV3 was applied to classify 52  
otolith images of red mullet (*Mullus barbatus*) (Politikos et al., 2021). While some 53  
state-of-the-art CNNs grew in model size a recent CNN architecture called EfficientNet 54  
(Tan and Le, 2019)(Tan and Quoc, 2019) demonstrated that increased performance 55  
could be achieved with smaller model sizes (number of parameters) using a compound 56  
scaling method for network depth, width and image size, resulting in a family of seven 57  
different models with different sizes. This network has been successfully applied with 58  
transfer learning to analyse images of salmon scales (Vabø et al., 2021)(Vabø et al., 59  
2021). Recently an even more compute efficient family of model architecture is called 60  
EfficientNetV2 (Mingxing Tan and, 2021) has become part of the state-of-the-art CNNs 61  
and has been made available. 62

In this study, we develop a learning framework for automating the age estimation of 63  
Atlantic cod based on multi-exposure images of broken otoliths. We apply the two 64  
EfficientNet family architectures EfficientNetV1 and EfficientNetV2 using three and two 65  
different model sizes from each family respectively. We compare the performance of the 66  
different models and discuss the use of ensemble of models to improve estimation 67  
accuracy. 68

## Method and materials 69

### Data Collection 70

We used a data set sampled from 5150 cod otoliths which was collected on surveys in 71  
the period 2012-2018 conducted by Institute of Marine Research (IMR) and aged by 72  
otolith experts. On each of the surveys, the otoliths were sampled using a 73  
random-stratified sampling based on fish length for each trawl station, and the otoliths 74  
from individual fish were randomly sampled. 75

The details of how the data-set was collected and sampled from surveys, camera and 76  
mount setup, how the otolith was processed before imaging, the resulting exposures, 77

and naming and folders organization can be found in (et al. et al., 2019) as well as  
where the data-set is available (<https://doi.org/10.21335/NMDC-1826273218>).  
78  
79

The 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°. We used the  
first 3 images, which positioned the otolith so the proximal surface was close to the top  
of the image. Figure 1 shows an example of the six image exposures taken of an otolith.  
80  
81  
82  
83

**Figure 1.** Otolith from 2016 with read age 6 years and light exposure medium, low,  
and high, then rotated 180° and three new images.  
84



The images were taken with a resolution of 3744×5616 pixels. The image light  
exposure varied depending on light condition outside, and was stored in the metadata of  
the JPG file. Typically the exposure order was middle-dark-light, then the rotation, and  
then middle-light-dark again, but the order could vary. The exact order was recovered  
by reading the 'ExposureTime' metadata property.  
85  
86  
87  
88

Figure 2 shows the age distribution of the 5150 otoliths in the data set, and figure 3  
shows the age distribution selected at random from the data set as the test set  
consisting of 10% of the whole data set (515 otoliths).  
89  
90  
91

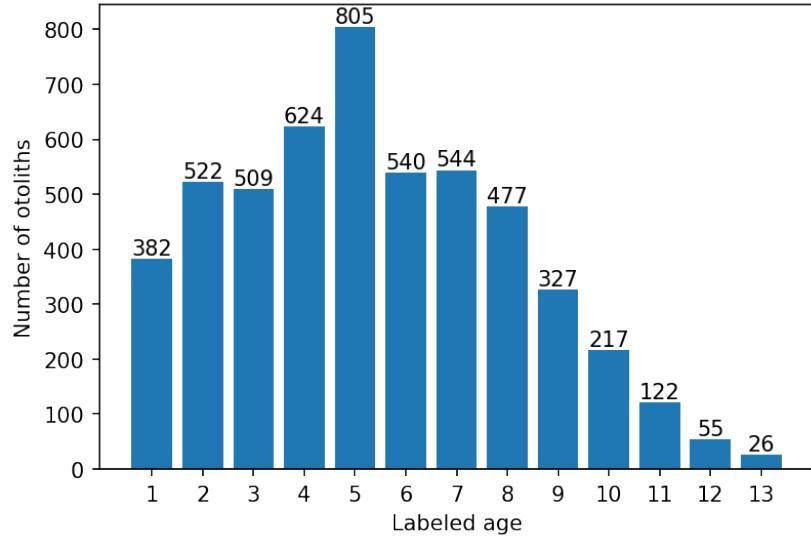
## Convolutional neural network architecture

  
92

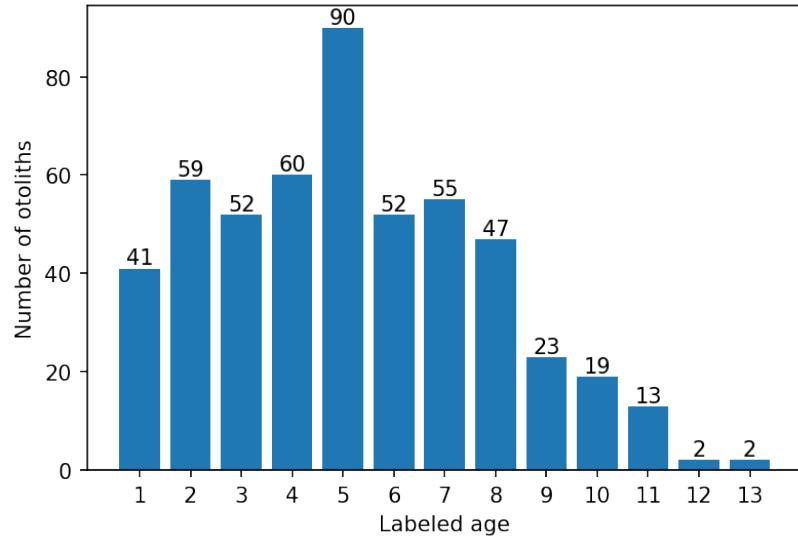
**Table 1.** EfficientNet and EfficientNetV2 models trained with image exposure. The  
models are numbered for reference of chapter on ensembles later  
93

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
All (3 images)	-	-	-	16	17

Each CNN was trained using transfer learning by loading ImageNet weights. The  
images were resized from 3744×5616 pixels to between 380×380 and 528×528 pixels  
93  
94



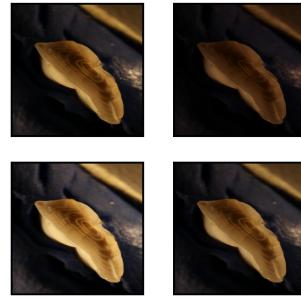
**Figure 2.** Age distribution of all 5150 images



**Figure 3.** Age distribution of 515 images from the test set

depending on the architecture. The pixel values have a range between 0 and 255, which  
95 was normalized to between 0 and 1. While test set size prediction was done on  $380 \times 380$   
96 and  $384 \times 384$  pixels. To investigate the image-taking protocol described in (et al. et al.,  
97 2019) we also trained on 9-channel images by stacking 3 RGB images representing 3  
98 different lighting exposures. Using Timm(Wightman, 2019), the imageNet weights were  
99 duplicated on the input layer to accommodate 9 channels. The three images used were  
100 of dark, medium and light exposure of the first orientation. Figure 4 shows an example  
101 of the 4 exposures used for training and testing the models.  
102

**Figure 4.** Otolith from 2013, read age: 6 years, and with light exposure: medium, low, high, and expectation per channel of the three exposures (9-channels).



CNNs were selected based on performance on the ImageNet benchmark and 103 availability of open-source implementations with imageNet weights. The imageNet 104 benchmark is for classification while we treated aging as a regression problem (Moen 105 et al., 2018) (Vabø et al., 2021). The last layer of the CNNs was modified to output a 106 linear output. In the EfficientNetV2 family we did this by applying three multi-layer 107 perceptron layers going from 1280 output of the last hidden layer to a dense 256-layer, 108 then a leakyRelu (Xu et al., 2015) layer, then a dense 32-layer, then a leakyRelu layer, 109 and finally a linear output layer. For EfficientNet we only changed the last layer from 110 softmax output to a linear output. 111

To each fold we normalized the age on the training-set by subtract the mean and 112 scaling to unit variance. The normalization was then applied to the validation and test 113 set. Test set predictions were obtained by applying the inverse transform. 114

## Implementation and training 115

EfficientNetV1 B4, B5, and B6 were imported and modified with TensorFlow (Abadi 116 et al., 2016) and Keras (Chollet and others, 2018) software packages in Python. 117 Computation was done using CUDA 11.1 and CuDNN with Nvidia(Nvidia Corp., Santa 118 Clara, California) A6000 accelerator card with 48 GB of GPU memory and P100 cards 119 with 12 GB of GPU memory, EfficientNetV2 Medium, and Large were imported and 120 modified with the PyTorch (Paszke et al., 2019) and Timm (Wightman, 2019) software 121 packages. Computation was done on P100 and RTX 3090 with 24 GB of GPU memory. 122 Pretrained weights for EfficientNet were available from Keras, and pretrained weights 123 for EfficientNetV2 were available from Timm. 124

Augmentation was applied to the training-set. The images were augmented using  
125 rotation between 0 and 360 degrees, and reflection by the vertical axis.  
126

The cost-function used was mean squared error (MSE) while the metric used for  
127 evaluating the models and comparing it to expert readers was accuracy. Accuracy was  
128 obtained by rounding the floating point number predictions to nearest integer and  
129 comparing the age classification against the true labels.  
130

The dataset of 5150 otoliths were divided into a training set constituting 90% of the  
131 otolith images (4635 otoliths) and a test set of 10% (515 otoliths). To get the most out  
132 of a small data-set we applied 10-fold cross-validation on the training set. This meant  
133 that 10% of the training set were used for validation and 90% (81% of the whole data  
134 set) were used for the actual training for each fold. Consequently 10 different models  
135 were trained with a different set of 463 images used for validation in each fold, i.e. each  
136 data point participates in the validation set once and in the training set 9 times.  
137 Among the 10 fold models, the one with the best MSE was chosen. The best  
138 model-parameters on the validation set were then used to predict the age on the test-set,  
139 and the metric for accuracy and MSE were recorded. The test-set is chosen at random,  
140 while the 10-fold split is chosen using stratified-kfold split, which preserves a similar  
141 distribution of the whole cross-validation set in each validation set. That means the 463  
142 images in the validation-set will have similar age distribution to that of the 4635 images  
143 in the cross-validation set.  
144

## Hyper-parameters

The CNN hyper-parameters configurations varied a little between the two families of  
146 networks, but were kept the same within the families. Some hyper-parameters that were  
147 tuned are batch size, learning rate, k-fold size, weight decay, step size, number of epochs,  
148 early stopping, and patience. Some parameters are constrained by the GPU memory,  
149 like batch-size which was kept at 8, except for the B6 model, which was run on the  
150 A6000 card.  
151

EfficientNet used learning-rate with weight decay scheduler, while EfficientNetV2  
152 used Cosine Annealing scheduler (Loshchilov and Hutter, 2016). The training- and  
153 validation image size used was as described in the papers, except for Large which uses  
154

smaller validation image size. The exact configuration of each network is available with  
 each network result in the GitHub page of the project  
<https://github.com/emoen/Deep-learning-for-regression-of-cod-otoliths>.  
155  
156  
157

**Table 2.** 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 3.** 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.

## Ensemble learning with averaging

158

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.  
159  
160  
161

There are many algorithms that perform ensembles to reach a prediction. E.g. bagging, stacking and boosting ensembles. We use simple ensemble average which is a form of bagging ensemble. Another example of a bagging ensemble is a voting ensemble.  
162  
163  
164

The ensemble average reduces the variance of the prediction and does not change the  
165

mean. The reduced variance improves the model performance. An ensemble can make  
166 better predictions and achieve better performance than any single contributing model,  
167 just as more experts will produce higher accuracy in predicting a single otolith. The  
168 ensemble prediction is therefor more robust because it reduces the spread of the  
169 predictions and model performance.  
170

We evaluate two types of simple ensemble average. The first ensemble is the average  
171 of the 10-fold cross-validation, which was reported as the model performance. This  
172 ensemble was reported as the performance of one model but we obtain 10 different  
173 models which contains the weights that gave the best MSE on the validation set, and  
174 the average of the prediction on the test set was reported as the accuracy after rounding.  
175

The second ensemble was created by combining models where we look at  
176 tuple-ensembles, consisting of 2 models, triplets, quadruples and so on, to ensemble of  
177 all 17 models which contained 20, 30 and so on up to 170 predictions on the test set.  
178 The accuracy was reported after rounding.  
179

By choosing the best model we were over fitting to the test set, but a subset of the  
180 best simple ensemble average learners will likely produce a better prediction on a  
181 hold-out test set than any of models.  
182

## Correlation of predictions on the test set and clustering analysis 183

We have looked at the correlations of predictions on the test set by creating a  
184 correlation matrix of each models prediction of each age class. This showed how much  
185 the models were in agreement with each other. Clustering analysis identified which  
186 models were more in agreement with each other.  
187

The relations between the model-predictions was linear therefor we used Pearson's  
188 correlation coefficient. Clustering analysis on the Pearson's correlation coefficient  
189 matrix was done by inspecting hierarchical clustering (HCA), and K-Means clustering  
190 with the number of clusters given by the elbow-, and the silhouette-score-method. With  
191 HCA we used Euclidean distance (Chebyshev distance, and Minkowski distance gave  
192 similar results), and we used Complete-Linkage clustering. The resulting clusters were  
193 drawn using a dendrogram. The key to interpreting the dendrogram is to focus on the  
194 height at which any two objects are joined together. The lower the height the more  
195

similar the two objects are.

196

## Results

197

The mean accuracy of the 17 models was 72.7% (table 4) on the test-set, and the  
198 standard deviation was 1.1. The least accurate model was B4-max, and the most  
199 accurate model was B5-min and B6-middle with accuracy of 74.4%. Assuming a normal  
200 distribution, the probability of seeing a model with lower accuracy than B4-max is less  
201 than 4.8% and the probability of seeing a model with higher accuracy than B5-min or  
202 B6-middle is less than 6.5%. The accuracy is not significantly different ( $p=0.05$ )  
203 between B5-min (or B6-middle) and B4-max model.  
204

B5 was the best model on all the exposures(min, middle, max) with a mean accuracy  
205 of 73.7%, and min-exposure was the best exposure with a mean accuracy of 73.3% Both  
206 B5 and B6 from the EfficientNet family was better than Medium and Large from the  
207 EfficientNetV2 family.  
208

**Table 4.** Mean accuracy on the test-set by light exposure and CNN architectures

Acc:light/CNN	B4	B5	B6	Medium	Large	Mean
min	72.8*	<b>74.4</b>	73.4*	74.0	72.0	73.3
middle	71.5	73.4	<b>74.4</b>	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

The mean MSE of the 17 models was 0.284 (table 5) on the test-set, and the  
209 standard deviation was 0.022. The highest MSE was from B5-max with MSE of 0.359,  
210 and the lowest MSE was from B6-middle exposure with MSE of 0.262. Assuming a  
211 normal distribution, then the probability of seeing a model with higher MSE than  
212 B6-max is less than 0.03% and the probability of seeing a model with lower MSE than  
213 B6-middle is less than 15.7%.  
214

Medium and Large were the best models with a MSE of 0.278, and the all-exposure  
215 (9-channel images) was the best exposures with a MSE of 0.272. The high MSE for  
216 B5-max and B6-max was due to a large misprediction of image with index 308 in the  
217 test-set (see chapter on Outliers).  
218

Table 6 shows percentage agreement (PA) between the models. Medium-all is the  
219 best model with PA 91.3% and B4-max is the worst model with PA 87.6%. Medium is  
220

**Table 5.** Mean MSE on the test-set by light exposure and CNN architectures

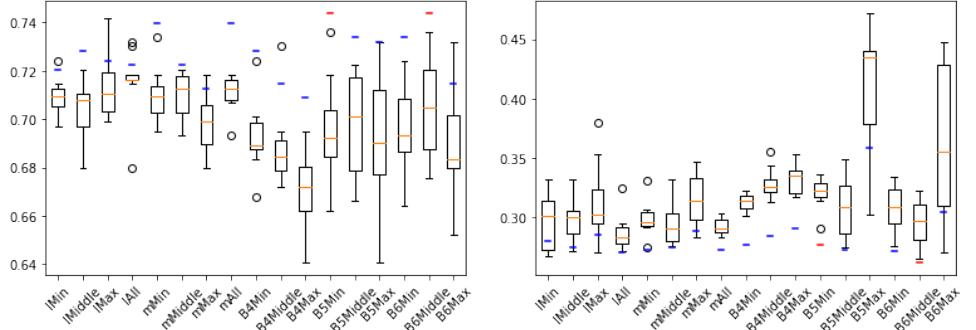
MSE:light/CNN	B4	B5	B6	Medium	Large	Mean
min	.277	.277	.272	.273	.280	.276
middle	.285	.273	<b>.262</b>	.278	.275	.275
max	.291	.359	.305	.289	.286	.306
9 channels	-	-	-	.273	.271	.272
Mean	.284	.303	.280	.278	.278	.284

the overall best model and B4 the worst. The all-exposure is the best exposure, and  
221  
max is the worst exposure.  
222

**Table 6.** Percentage Agreement between models

PA:light/CNN	B4	B5	B6	Medium	Large	Mean
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	-	-	-	<b>91.3</b>	91.1	91.2
Mean	88.1	89.8	89.0	90.4	89.8	89.6

Figure 5 shows a box-plot of each 10-fold ensemble average prediction on accuracy,  
223  
and MSE for all the 17 models. The red lines are the ensemble-average predictions with  
224  
highest accuracy. The blue lines are the other ensemble average predictions. The orange  
225  
lines are the mean accuracy or MSE. The ensemble metric was either better than or in  
226  
the upper quantile for all the models. The prediction MSE and accuracy of each fold are  
227  
given in table 13 and 14 in appendix C.  
228

**Figure 5.** A box-plot of accuracy score (left) and MSE (right) of all the 17 models and 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 accuracy. The orange lines are the mean of the 10-fold predictions.

## Prediction by age class

229

When taking the accuracy of all models by age class, we found that accuracy for one-  
and two-year-old's was better than 90% (figure 6). All age classes six years or younger  
were correctly classified with more than 70% accuracy, and all 13-year-old's were  
predicted to be younger (see Figure 15 in appendix B which shows model the mean and  
standard deviation from the residuals test set prediction by age classes).

230

231

232

233

234

No systematic bias in the age prediction of CNN is visible, except for the  
underestimated age of individuals aged by human reader as 13 year old (figure 7).

235

236

237

## Simple ensemble-average predictions

238

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 combination we  
recorded the best ensemble and found that the best overall ensemble-average 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.

239

240

241

242

243

Table 7 shows first the number of combinations of models that exists of tuples,  
triplets and so on labeled with 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  
1. Table 8 shows the same information but selected to minimize MSE.

244

245

246

247

248

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.

249

250

251

252

253

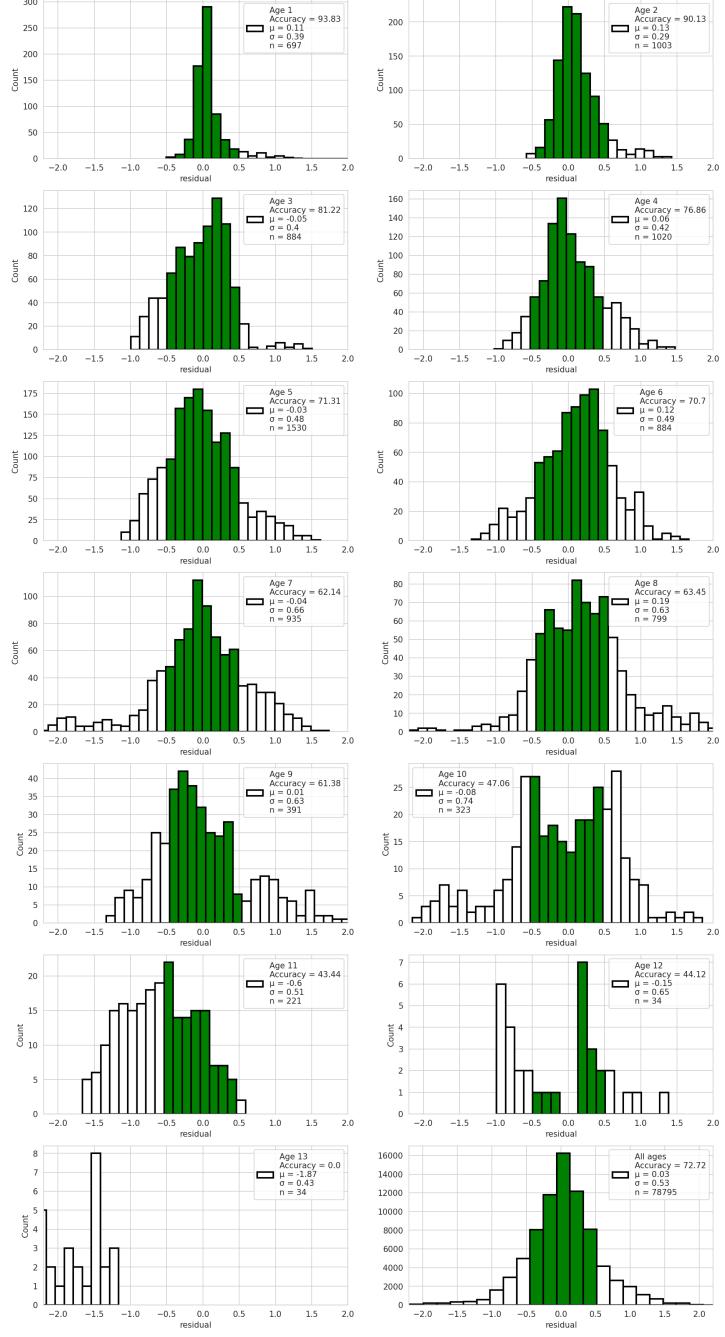
We observe that model B4-min (No 1), and B6-min (No 3) were the best models  
with inclusion in 14 ensembles(table 9). These models did not have the highest accuracy  
(B5-min, and B6-middle with 74.6%) but an accuracy of 72.8% and 73.4%. This was

254

255

256

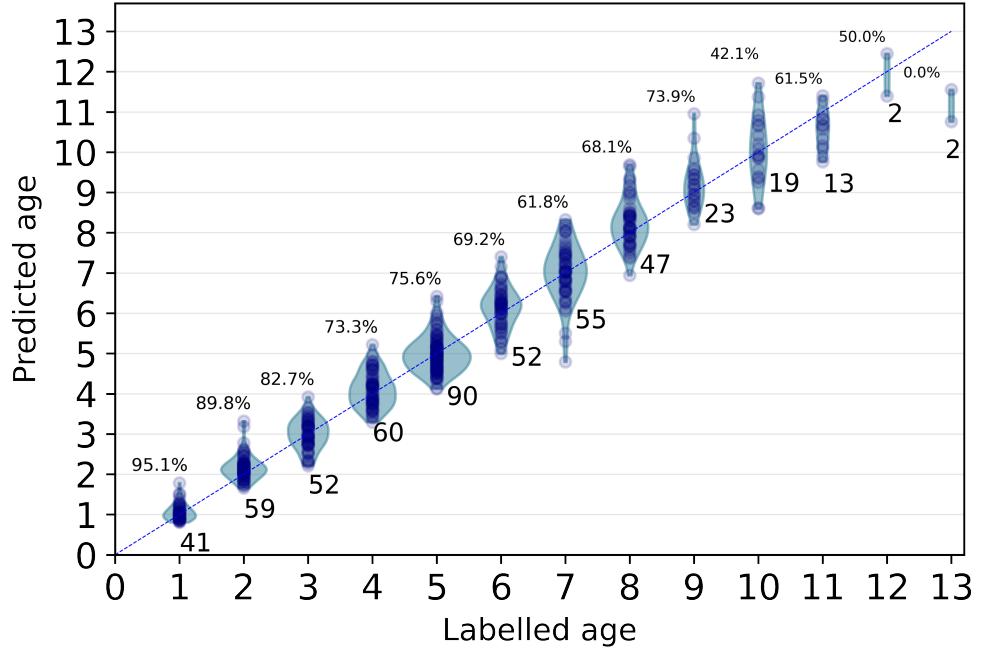
**Figure 6.** Predictions by age class from the average of all models. The green region shows the correctly classified age after rounding. The axis is fixed, hence large outliers will not be visible.



lower than the highest accuracy models, which was B5-min and B6-middle, which had a rank of 3 and 5 respectively. 257

Exposure-types ranked by how many times they were included in an ensemble was as follows: min-exposure (rank 4.4), middle-exposure (rank 8.6), all-exposures (rank 10), 258  
259  
260

**Figure 7.** 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 7.** Binomial combinations of simple average of ensembles accuracy

Coeff	#Comb	Best	Mean	Ensemble (see table 1)
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)

and max-exposure (rank 11.2). EfficientNet had a rank of 6.6 and EfficientNetV2 had a rank of 10.3.

261

262

Table 10 compares the accuracy of the best ensemble with the mean of all 17 models

264

**Table 8.** Binomial combinations of simple average of ensembles MSE

Coeff	#comb	best	Mean	Ensemble (see table 1)
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 9.** Rank of number of times a model is in an ensemble selected by 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

by age classes. The accuracy of the mean of the models was shown in figure 5. The  
 table shows that the best ensemble improved the accuracy of prediction for all age  
 classes, except 13-year-old's which had 0% accuracy. A distribution plot like figure 5 for  
 the best ensemble can be found in appendix D as figure 16.

265

266

267

268

269

**Table 10.** 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-old's.

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

## Outliers

Figure 8 shows 6 images which had an error after rounding of more than 1 year. All the images with more than 1 year in prediction error are shown in table 11, with comments by an expert on the most common mispredictions in table 12. In general, large outliers occurred throughout all of the tested models and ensembles. Most outliers were visually challenging images with artifacts and / or low readability. For example, image 13 was overestimated in all B models, likely due to a settlement ring in the inner core region that a human reader would 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 interrupt 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 misreported.

**Figure 8.** Images with index 13, 71, 270, 342, 362 and 369 from the test-set was miss-predicted by between 25% and 100% of the models



We observed that there were some cod otoliths that 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), to some models and on one exposure (E.g otolith

308), and to both family of models and on some exposures (E.g. otolith 320). 288

We also observed that the number of large outliers did not correlate with model 289 performance like B5-min, and B6-mid which had 7 and 9 outliers, but the best accuracy. 290 While B4-max with the lowest accuracy (70.9%) had the least number of large outliers 291 with only 6 mispredictions. 292

## Correlation of predictions and cluster analysis 293

The correlation of models on the test-set predictions given in figure 9 show that the 294 models correlates a lot on outlier predictions. The correlation from all the predictions 295 on the test set varied between 0.988 to 0.999, with the lowest correlation found between 296 B5-min and Medium-min. The correlation was calculated using Pearson's correlation 297 with Euclidean distance. 298

From figure 9 we saw that the EfficientNetV2 family was correlated from the red 299 block in the lower right corner. The dendrogram of figure 10 shows hierarchical 300 clustering (HCA) of the models. HCA found 3 clusters, b5-min, and B6-middle, which 301 are the two best models, a cluster of all the EfficientNetV2 models, and a cluster of the 302 rest of the models (figure 10). 303

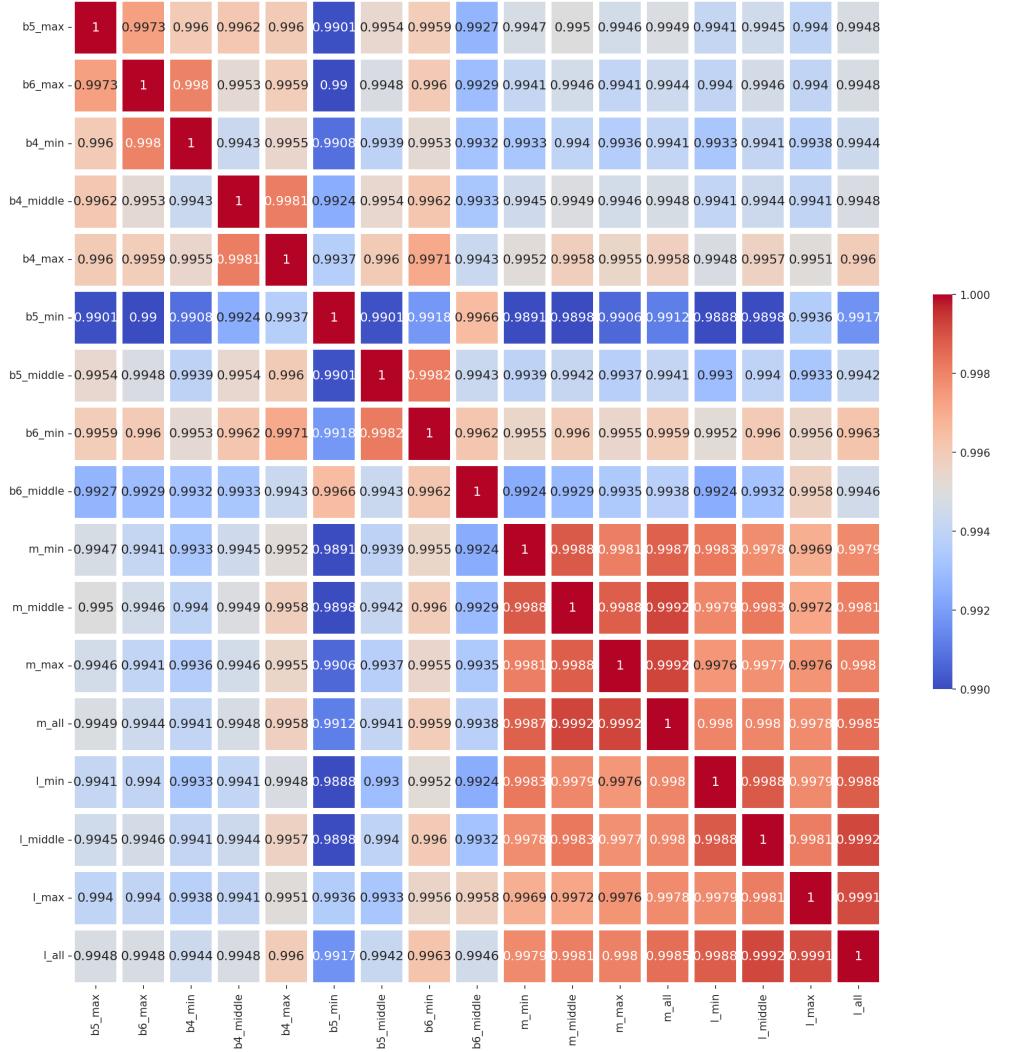
Using K-Means together with the elbow- and silhouette-score-methods to find the 304 optimal number of clusters, both found that there were 3 clusters in the correlation 305 matrix (figure 11). 306

With K-Means and using 3 clusters we found similar clusters to HCA with the 307 clusters: 308

- B5-max, and B6-max 309
- B4-min, B5-min, B6-min, B4-middle, B5-middle, B6-middle, B4-max 310
- EfficientNetV2 family of models 311

Figure 12 shows a scatter plot of two of the least correlated models, B5-min and 312 Medium-min, which had Pearson's correlation 0.988. The red point inside the two 313 circles is not a data point but "bullseye" if both models predict the correct age. The 314 last sub-figure was the residual correlation of all age classes. It can be viewed in more 315 details in figure 13. 316

**Figure 9.** Pearson correlation of each model prediction on the test-set

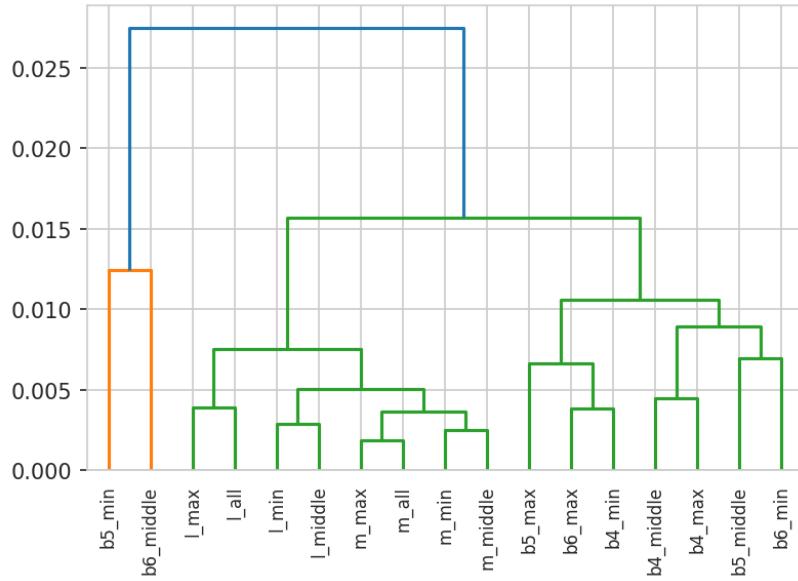


## Discussion

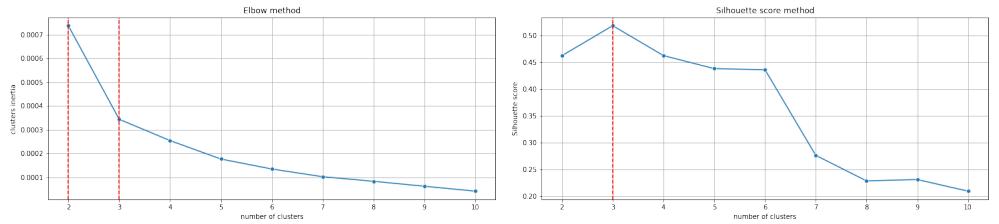
During initial training, we trained a B4 network on ca 2000 images and obtained an accuracy of ca 60%, later another 3000 images were added and the same network was trained on ca 5000 images which resulted in accuracy of ca 70%. It could be interesting to investigate if adding another 3-5000 images would increase the accuracy to 80%.

To reach human level accuracy a score of 85% or higher is required ref-needed, and a score of 90% is considered good. Figure 14 shows a sample of 25 predictions on the validation-set during training.

**Figure 10.** hierarchical clustering (HCA) on correlation of predictions



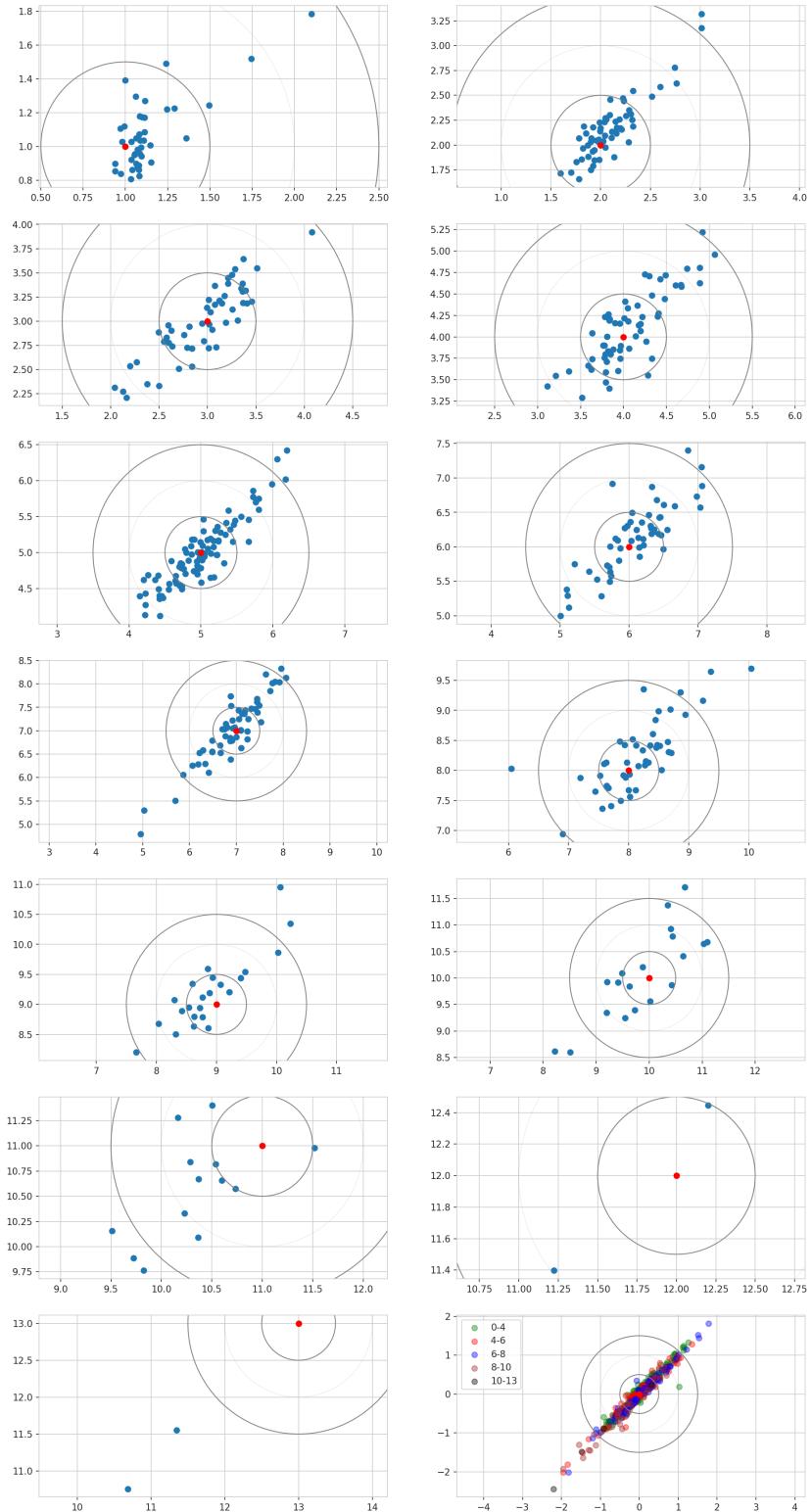
**Figure 11.** 1.Elbow method and 2. silhouette-score method



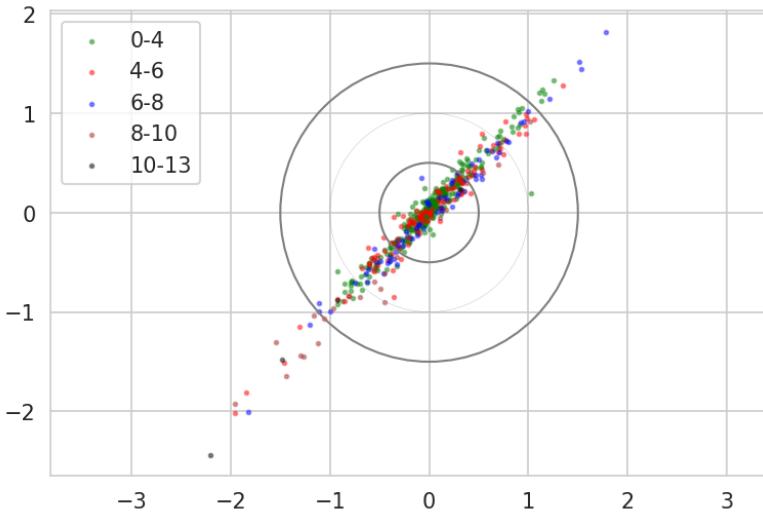
## The effect of data size

A crucial issue in machine learning projects is to determine how much training data is needed to achieve a specific performance goal. In computer vision, one commonly used rule of thumb adopted from the number of images and classes in the ImageNet dataset, is to have a thousand images for each class. In our case of cod otolith images the task entails regression towards 13 age classes instead of classification into 1000 classes. Therefore, approximately 13,000 images would appear to be the optimal number for our problem based on the rule of thumb for computer vision. This number can be reduced if transfer learning is applied for images within a similar domain. For cod otolith images, the domain is different than images in ImageNet. However, despite the different image domain for our problem we do see a significant performance boost in using transfer learning, suggesting that fewer images are needed than if trained from scratch. Excessive use of augmentation also reduces the number of images required. On the

**Figure 12.** Scatter plot of each age-class by Medium-min  $\times$  B5-min.



**Figure 13.** Scatter plot of the residuals of all age classes by Medium-min  $\times$  B5-min.



**Figure 14.** Sample of 25 predictions on a model of training on EfficientNetV2 size medium with minimum light exposure, left number is prediction, and right number is age read

other hand, a general insight from deep learning is that more training data is always  
338  
advantageous. Given the use of transfer learning and augmentation, the number of  
339  
images used in this study, around 5000, might be close to the optimal, but we still think  
340  
that a larger training set would improve performance. During initial training we trained  
341

a B4 network on about 2000 images and obtained an accuracy of around 60%. Later  
342 another 3000 images was added and the same network was trained on around 5000  
343 images which resulted in accuracy of about 70%. This suggests to us that if our training  
344 set where even larger, say 10,000 images, this would boost performance further, maybe  
345 even approaching human level accuracy of 85%.  
346

## Accuracy for different age classes

All models tend to predict younger year classes with greater accuracy than older year  
348 classes. Pooling all models, prediction accuracy for 1–3 year old otoliths is ~80%,  
349 exceeding 90% for one year old's. Accuracy tends to decline with increased age,  
350 especially for otoliths older than 9 years for which the training set is less represented  
351 than for younger otoliths (Figure 2). On the other hand, age 5 is the most abundantly  
352 represented age class in the training set, and accuracy for this age class is lower than  
353 the less abundant one-year-old age class. Thus, the CNN appears to be particularly  
354 competent at analyzing cod otoliths with fewer growth zones. . . Let us discuss why  
355 this is so... Ask the readers/biologists.  
356

## Ensemble of models

We should discuss the improved accuracy gained when combining models into ensemble  
358 of models (table 8). The best combinations show accuracy greater than 75%. Add a  
359 comment on the relation between variation in accuracy and the resulting accuracy of  
360 the ensemble (predictions of B4/5/6 have higher variance, but their ensembles win).  
361

## Moved from introduction to Discussion –

Accuracy, efficiency and cost benefits of CNN classifiers versus manual reading  
363 . . . Despite fast progress the results remain mixed and often yield lower precision and  
364 consistency than those obtained by trained human readers, which limits the application  
365 of automated methods in real conditions. However, one aspect that is often under  
366 considered by such studies are the practical time and cost benefits that implementing a  
367 functional ML framework would provide. As noted by Fisher and Hunter (2018) (Fisher  
368 and Hunter, 2018) in their review of digital techniques for otolith analysis, “costs for  
369

human and machine ageing systems are broadly similar since a large part of the cost is  
370 associated with preparing the otolith sections”. As such, the net benefit of automated  
371 ageing routines is directly dependent on the ability to scale performance using a  
372 comparatively smaller number of samples than human readers or, alternatively, to train  
373 them on “rougher” data that can be produced faster and at a more efficient cost. Also,  
374 CNN can be applied without high additional cost or even be incorporated in the routine  
375 protocols, but add a new value e.g., reading consistency check, time-drifts evaluations,  
376 inter-reader comparisons (how much ‘off’ is each reader when compared to the CNN  
377 predictions, even if not compared with the same otolith samples), etc.  
378

We see the process of CNN implementation as an evolution of the protocols, with  
379 the intensive phase of model development and training. Gradual improvement of model  
380 reliability should then allow for the application of CNN as a complementary supportive  
381 tool for the age traditional estimations. Finally, this change should aim to scale the  
382 capacity of the age reading experts and improve sampling in the areas, fish stocks, or  
383 periods that lack proper reading effort.  
384

## Discussion by Côme

The different networks were generally good at interpreting otoliths ages within one year  
385 of human readers. One common issue seemed to be with otoliths of poor readability, in  
386 particular drastic changes in exposure or visual damages and interruptions on the  
387 reading plane. This is likely due to how the networks utilize the visual information to  
388 interpret ages: human readers will generally divide the otolith in three axes (dorsal,  
389 ventral, distal) and will count rings within a limited area along those axes, using all  
390 three to double-check as needed when a specific axis is partially unreadable or in case of  
391 doubt. Alternatively, the current machine learning framework does not provide visual  
392 clues of where the estimation of age is made, and it is therefore possible that a specific  
393 axis carrying the most information is usually chosen by the network to make an  
394 estimation. In cases where this axis is harder to read or damaged, the algorithm may  
395 therefore struggle to verify its estimate using the rest of the image. A future  
396 improvement would be to add visual feedback of where the estimates are made, and  
397 ideally a dynamic system where human readers could accept or reject individual rings  
398

identified by the algorithm to further train it. In addition, the current implementation  
400  
of the algorithm requires a significant compression of the images in order to reduce  
401 computing time and costs. The compressed images therefore carry their information in  
402 a reduced number of pixels that may create issues for difficult to read areas. This is  
403 especially true due to the somewhat curved or oval shape of the otolith, as a compressed  
404 image will not only have less pixels to work with but will also have a comparatively  
405 more important fraction of black background, which is effectively useless for age  
406 interpretation. Improvements could therefore be made by instead first isolate the  
407 otoliths from their background, in order to have a machine learning network trained on  
408 using exclusively the information contained within the area of interest.  
409

Need discussion on the “value” of this work from a biology point of view. As stated  
410 in introduction: an efficient network usable on data collected at sea (imaging is easy)  
411 means we could save a lot of resources and time. How to improve it: needs transparency  
412 so that we know it reads the right things. Ideally, a network that gives you visual clues  
413 of where and how the age is estimated. For example, it could be entirely plausible to  
414 develop a network that has a very high accuracy compared to human readers, but  
415 actually does not guess ages based on actual rings like we hoped to do but simply by  
416 looking at size ratios. Need for more biological breadth: for now we trained it exclusively  
417 on NEA cod otoliths, but ideally we would need something that can read and interpret  
418 different cod populations with their own ring pattern instead of a one-fits-all algorithm  
419

I suggest having a chat with Szymon, he and I will be on the same page for all these  
420 biological aspects of otolith reading so he should be able to contribute here.  
421

## Conclusion 422

Our results demonstrate that the use of deep learning techniques in the analysis of  
423 otoliths have a major potential for facilitating automation. We believe that carefully  
424 trained CNNs could become a major component in automated pipelines that require  
425 minimal processing and could be able to produce near at sea age estimates.  
426

When developing the framework for the automatic age estimation, it is advised to  
427 include B4 architectures as they are quick to train, and performs good. Ensemble  
428 approaches are also recommended if more effort is favorable, as it gives a more robust  
429 and higher performing prediction. For a quick-to-train ensemble, B5 and Medium could  
430

be added. It is recommended to use under-exposed images.

431

## References

432

### References

433

- 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>.  
et al., M., Thorsen, A., Godiksen, J., Malde, K., and Handegard, N. (2019). An efficient  
protocol and data set for automated otolith image analysis. *GeoScience Data Journal*.  
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.  
<https://cdnsciencepub.com/doi/abs/10.1139/f04-063> (Accessed 3 February 2022).
- 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. *Marine Ecology Progress Series*, 426: 1–12.
- 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. *Fisheries Research*, 96: 319–322.
- 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, *Advances in Neural Information Processing Systems* 25, pages 1097–1105. Curran Associates, Inc.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *neurips*.
- Mahé, K. (2009). Project no. 044132. *Automated FISH Ageing (AFISA): final activity report*.
- Mingxing Tan and, Q. V. L. (2021). Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298.
- Moen, E., Handegard, N. O., Allken, V., Albert, O. T., Harbitz, A., and Malde, K. (2018). Automatic interpretation of otoliths using deep learning. *Plos One*.
- Panfili, J., de Pontual, H., Troadec, H., and Wrigh, P. J. (2002). Manual of fish sclerochronology. <https://archimer.ifremer.fr/doc/00017/12801/> (Accessed 3 February 2022).

- 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. 485
- 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. 492
- 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. ””. 495
- 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. 498
- 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. 501
- Siskey, M. R., Wilberg, M. J., Allman, R. J., Barnett, B. K., and Secor, D. H. (2016). Forty years of fishing: changes in age structure and stock mixing in northwestern atlantic bluefin tuna (*thunnus thynnus*) associated with size-selective and long-term exploitation. *ICES Journal of Marine Science*, 73: 2518–2528. 504
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567. 508
- Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946. 510
- Tyler, A. V., Beamish, R. J., and McFarlane, G. A. (1989). Implications of age determination errors to yield estimates. *ICES Journal of Marine Science*, 108: 27–35. 512

- Vabø, R., Moen, E., Smoliński, S., Åse Husebø, Handegard, N. O., and Malde, K. 514  
(2021). Automatic interpretation of salmon scales using deep learning. *Ecol. Inform.* 515  
63, 101322 (2021). 516
- Wightman, R. (2019). Pytorch image models. 517  
<https://github.com/rwightman/pytorch-image-models>. 518
- Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified 519  
activations in convolutional network. *CoRR*, abs/1505.00853. 520

## A Prediction error of more than 1.5 year

521

**Table 11.** Predictions error with residual of more than 1.5 year 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					10.7					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								9.4	
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									9.4	
m-min			5.0	11.3		10.0					10.7	5.0	8.2										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										9	
m-all			5.0	11.2		10.1					10.5	5.3	8.2										8	
l-min			5.1	11.5		9.8	9.3				10.7	5.2	8.3										8	
l-mid			5.0		9.8	9.4	5.5				10.6	5.2	8.1	10.5									9	
l-max			9.5	5.1		9.9	3.6	5.4			10.8	5.1	8.2										10	
l-all			9.3	5.0		9.8					10.8	5.2	8.0	10.5									9	
Age	8	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	1	17	17	6	2	7	2	1	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	-	

522

**Table 12.** Comments on the most frequently miss predicted 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 very bright on all three axis, and the dorsal axis had a break line, 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 of good quality, but there was 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 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.
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 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 over-exposed.

524

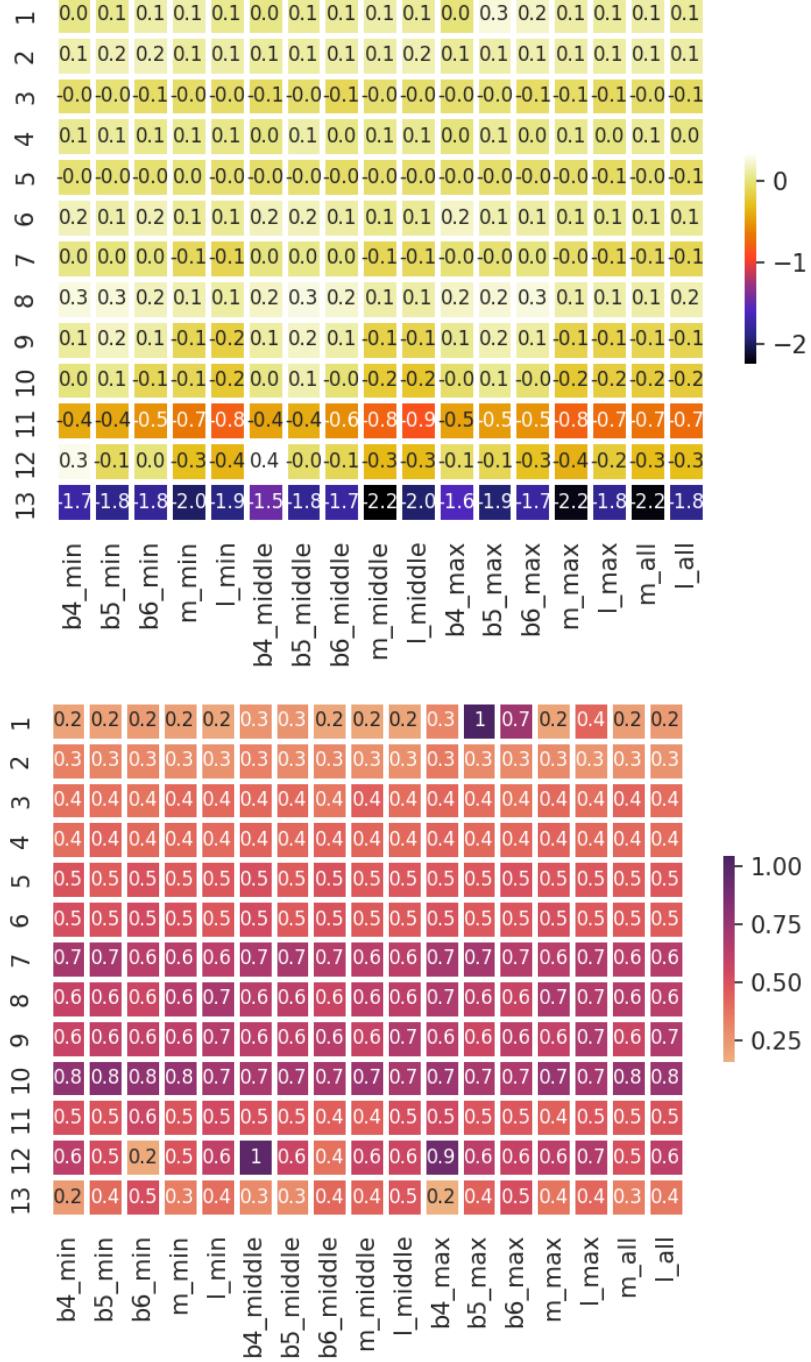
## B Model mean and standard deviation of residual

525

### test set prediction per age class

526

**Figure 15.** Model mean and standard deviation of residual test set prediction by age class



## C Model accuracy and MSE per fold

527

**Table 13.** 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

528

**Table 14.** 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

## D Prediction per age class using from best ensemble

530

**Figure 16.** Predictions by age class from the best ensemble.

