

Automatic interpretation of cod otoliths using deep learning

Endre Moen^{1*}, Rune Vabø¹, Szymon Smoliński³, Come Denechaud¹, Ketil Malde^{1,2},

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).
Therefore, streamlining, scaling, and increasing the quality of age estimations can
improve the reliability of evaluations of fish biology and consequently stock size and
structure (Beamish and McFarlane, 1995; Ragonese, 2018; Tyler et al., 1989).

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 otoliths that typically require time-consuming preparation.
These routines vary between populations and institutes and range from direct reading of
broken otoliths under a magnifying glass, to embedding, thin sectioning and finally
imaging of the sections under a microscope. There has been a variety of methods
proposed to automatically interpret otoliths, which range from one-dimensional data
analysis like intensity transects (Mahé, 2009) to the more recent effort toward
developing machine learning (ML) frameworks (Moen et al., 2018; Politikos et al., 2021).

About deep learning and image analysis

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

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

Convolutional Neural networks in aging otoliths 63

There are two families of models used, EfficientNet with CNNs B0-B7 (Tan and Le, 64
2019) and EfficientNetV2 with convolutional neural-networks (CNNs) small, medium, 65
Large, and Xtra-Large (Tan and Le, 2019). The EfficientNet family of models, was 66
introduced in 2019 and the largest model B7 achieved state-of-the-art result on the 67
ImageNet (Deng et al., 2009) benchmark. It uses neural architecture search to scale 68
image-size and the network. The EfficientNetV2 family of models was introduced in 69
2021 and Xtra-large achieved state-of-the-art result on the ImageNet benchmark again. 70
It extends on the previous work and introduces new ideas, like scaling up test-set 71
image-size. In this work we investigate EfficientNet B4-B6, and EfficientNetV2 medium 72
and large which shows the best compromise between training-time and accuracy. 73

In this study, we develop a learning framework for automating the age estimation of 74
Atlantic cod based on multi-exposure images of broken otoliths. We apply the two 75
EfficientNet family architectures EfficientNetV1 and EfficientNetV2 using three and two 76
different model sizes from each family respectively. We compare the performance of the 77
different models and discuss the use of ensemble of models to improve estimation 78

accuracy.

79

Method and materials

80

Data Collection

81

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

82

83

84

85

86

87

88

89

90

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

87

88

89

90

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.

91

92

93

94

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



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.

95

96

97

98

99

Figure 2 shows the age distribution of the 5150 otoliths in the data set, and 3 shows
100 the age distribution selected at random from the data set used for testing and consisting
101 of 10% of the whole data set or 515 otoliths.
102

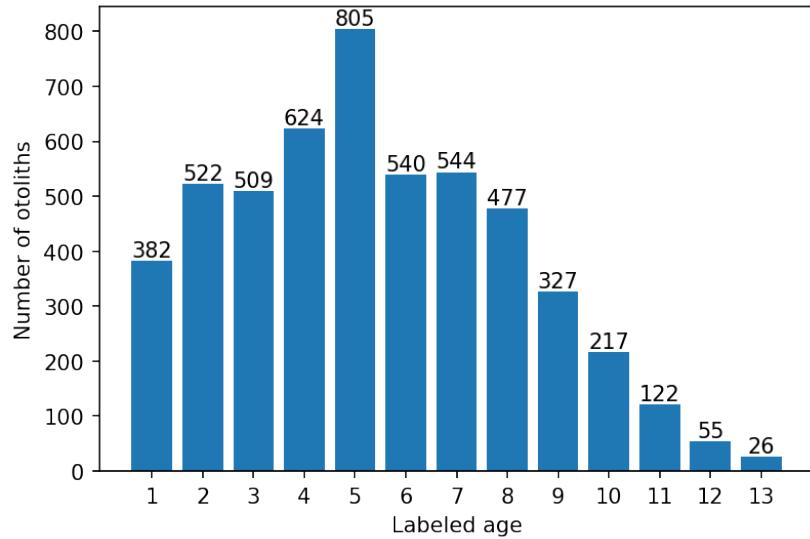


Figure 2. Age distribution of all 5150 images

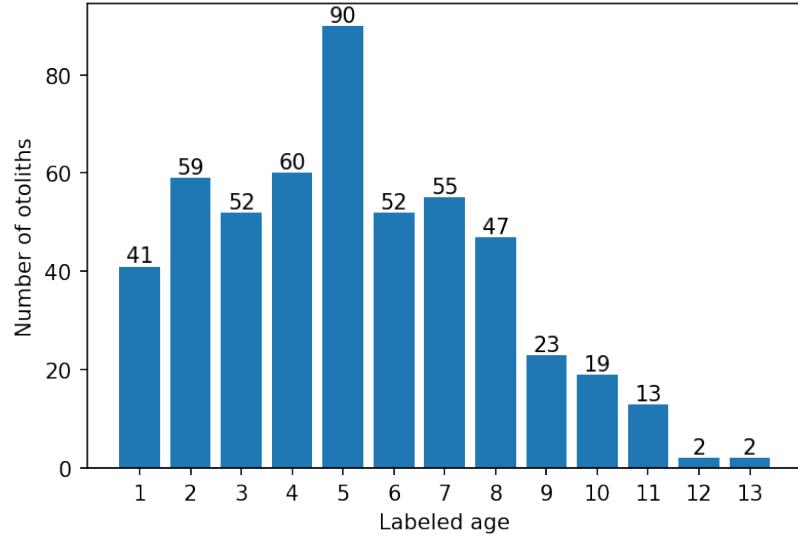


Figure 3. Age distribution of 515 images from the test set

Convolutional neural network architecture

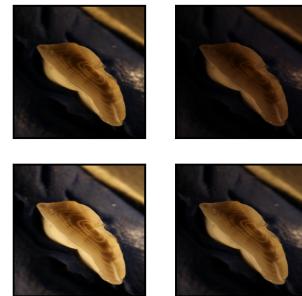
Each CNN was trained using transfer learning by loading ImageNet weights. The
103 images were resized from 3744×5616 pixels to between 380×380 and 528×528 pixels
104
105

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

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

depending on the architecture. The pixel values have a range between 0 and 255, which
106 was normalized to between 0 and 1. While test set size prediction was done on 380×380
107 and 384×384 pixels. To investigate the image-taking protocol described in (et al. et al.,
108 2019) we also trained on 9-channel images by stacking 3 RGB images representing 3
109 different lighting exposures. Using Timm(Wightman, 2019), the imageNet weights were
110 duplicated on the input layer to accommodate 9 channels. The three images used were
111 of dark, medium and light exposure of the first orientation. Figure 4 shows an example
112 of the 4 exposures used for training and testing the models.
113

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



CNNs were selected based on performance on the ImageNet benchmark and
114 availability of open-source implementations with imageNet weights. The imageNet
115 benchmark is for classification while we treated aging as a regression problem (Moen
116 et al., 2018) (Vabø et al., 2021). The last layer of the CNNs was modified to output a
117 linear output. In the EfficientNetV2 family we did this by applying three multi-layer
118 perceptron layers going from 1280 output of the last hidden layer to a dense 256-layer,
119 then a leakyRelu (Xu et al., 2015) layer, then a dense 32-layer, then a leakyRelu layer,
120 and finally a linear output layer. For EfficientNet we only changed the last layer from
121 softmax output to a linear output.
122

To each fold we normalized the age on the training-set by subtract the mean and
123 scaling to unit variance. The normalization was then applied to the validation and test
124 set. Test set predictions were obtained by applying the inverse transform.
125

Implementation and training

126

EfficientNetV1 B4, B5, and B6 were imported and modified with TensorFlow (Abadi
127 et al., 2016) and Keras (Chollet and others, 2018) software packages in Python.
128 Computation was done using CUDA 11.1 and CuDNN with Nvidia(Nvidia Corp., Santa
129 Clara, California) A6000 accelerator card with 48 GB of GPU memory and P100 cards
130 with 12 GB of GPU memory, EfficientNetV2 Medium, and Large were imported and
131 modified with the PyTorch (Paszke et al., 2019) and Timm (Wightman, 2019) software
132 packages. Computation was done on P100 and RTX 3090 with 24 GB of GPU memory.
133 Pretrained weights for EfficientNet were available from Keras, and pretrained weights
134 for EfficientNetV2 were available from Timm.
135

Augmentation was applied to the training-set. The images were augmented using
136 rotation between 0 and 360 degrees, and reflection by the vertical axis.
137

The cost-function used was mean squared error (MSE) while the metric used for
138 evaluating the models and comparing it to expert readers was accuracy. Accuracy was
139 obtained by rounding the floating point number predictions to nearest integer and
140 comparing the age classification against the true labels.
141

The dataset of 5150 otoliths were divided into a training set constituting 90% of the
142 otolith images (4635 otoliths) and a test set of 10% (515 otoliths). To get the most out
143 of a small data-set we applied 10-fold cross-validation on the training set. This meant
144 that 10% of the training set were used for validation and 90% (81% of the whole data
145 set) were used for the actual training for each fold. Consequently 10 different models
146 were trained with a different set of 463 images used for validation in each fold, i.e. each
147 data point participates in the validation set once and in the training set 9 times.
148

Among the 10 fold models the one with the best MSE was chosen. The best
149 model-parameters on the validation set were then used to predict the age on the test-set,
150 and the metric for accuracy and MSE were recorded. The test-set is chosen at random,
151 while the 10-fold split is chosen using stratified-kfold split, which preserves a similar
152

distribution of the whole cross-validation set in each validation set. That means the 463
153 images in the validation-set will have similar age distribution to that of the 4635 images
154 in the cross-validation set.
155

Hyper-parameters

The CNN hyper-parameters configurations varied a little between the two families of
157 networks, but were kept the same within the families. Some hyper-parameters that were
158 tuned are batch size, learning rate, k-fold size, weight decay, step size, number of epochs,
159 early stopping, and patience. Some parameters are constrained by the GPU memory,
160 like batch-size which was kept at 8, except for the B6 model, which was run on the
161 A6000 card.
162

EfficientNet used learning-rate with weight decay scheduler, while EfficientNetV2
163 used Cosine Annealing scheduler (Loshchilov and Hutter, 2016). The training- and
164 validation image size used was as described in the papers, except for Large which uses
165 smaller validation image size. The exact configuration of each network is available with
166 each network result in the GitHub page of the project
167

(<https://github.com/emoen/Deep-learning-for-regression-of-cod-otoliths>).
168

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`

Ensemble learning with averaging

Ensemble learning is an algorithm that combines the predictions from multiple models
170 to reach a final prediction, and obtains a predictive performance that is better than any
171 of the constituent models alone.
172

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.

There are many algorithms that perform ensembles to reach a prediction. E.g. 173
bagging, stacking and boosting ensembles. We use simple ensemble average which is a 174
form of bagging ensemble. Another example of a bagging ensemble is a voting ensemble. 175

The ensemble average reduces the variance of the prediction and does not change the 176
mean. The reduced variance improves the model performance. An ensemble can make 177
better predictions and achieve better performance than any single contributing model, 178
just as more experts will produce higher accuracy in predicting a single otolith. The 179
ensemble prediction is therefore more robust because it reduces the spread of the 180
predictions and model performance. 181

We evaluate two types of simple ensemble average. The first ensemble is the average 182
of the 10-fold cross-validation, which was reported as the model performance. This 183
ensemble was reported as the performance of one model but we obtain 10 different 184
models which contains the weights that gave the best MSE on the validation set, and 185
the average of the prediction on the test set was reported as the accuracy after rounding. 186

The second ensemble was created by combining models where we look at 187
tuple-ensembles, consisting of 2 models, triplets, quadruples and so on, to ensemble of 188
all 17 models which contained 20, 30 and so on up to 170 predictions on the test set. 189
The accuracy was reported after rounding. 190

By choosing the best model we were over fitting to the test set, but a subset of the 191
best simple ensemble average learners will likely produce a better prediction on a 192
hold-out test set than any of models. 193

Correlation of predictions on the test set and clustering analysis 194

We have looked at the correlations of predictions on the test set by creating a 195 correlation matrix of each models prediction of each age class. This showed how much 196 the models were in agreement with each other. Clustering analysis identified which 197 models were more in agreement with each other. 198

We used Pearson's correlation coefficient as the relationship between the 199 model-predictions on the test-set was identified as linear. Clustering analysis was done 200 by inspecting hierarchical clustering (HCA) and K-Means clustering with the number of 201 clusters given by the elbow-, and the silhouette-score-method. With HCA we used 202 Euclidean distance (Chebyshev distance, and Minkowski distance gave similar results), 203 and we used Complete-Linkage clustering. The resulting clusters were drawn using a 204 dendrogram. The key to interpreting the dendrogram is to focus on the height at which 205 any two objects are joined together. The lower the height the more similar the two 206 objects are. 207

Results 208

Table 4 and 5 shows the accuracy and MSE we got from the 10-fold cross-validation 209 training as a simple ensemble average prediction on the test-set 210

The mean accuracy of the 17 models was 72.7% (table 4) on the test-set, and the 211 standard deviation was 1.1. The least accurate model was B4-max, and the most 212 accurate model was B5-min and B6-middle with accuracy of 74.4% Assuming a normal 213 distribution, then the probability of seeing a model with lower accuracy than B4-max is 214 less than 4.8% and the probability of seeing a model with higher accuracy than B5-min 215 or B6-middle is less than 6.5%. The accuracy is not significantly different ($p=0.05$) 216 between B5-min and B4-max model. 217

B5 was the best model on all the exposures with a mean accuracy of 73.7%, and 218 min-exposure was the best exposure with a mean accuracy of 73.3% Both B5 and B6 219 from the EfficientNet family was better than Medium and Large from the 220 EfficientNetV2 family. 221

The mean MSE of the 17 models was 0.284 (table 5) on the test-set, and the 222 standard deviation was 0.022. The highest MSE was from B5-max with MSE of 0.359, 223

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*	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

and the lowest MSE was from B6-middle exposure with MSE of 0.262. Assuming a
normal distribution, then the probability of seeing a model with higher MSE than
B6-max is less than 0.03% and the probability of seeing a model with lower MSE than
B6-middle is less than 15.7%.
224
225
226
227

Medium and Large was the best models with a MSE of 0.278, and the all-exposure
(9-channel images) was the best exposures with a MSE of 0.272. The high MSE for
B5-max and B6-max was due to a large missprediction of image with index 308 in the
test-set (see chapter on Outliers).
228
229
230
231

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	.262	.278	.275	.275
max	.291	.359	.305	.289	.286	.306
9 channels	-	-	-	.273	.271	.272
Mean	.284	.303	.280	.278	.278	.284

Figure 5 shows a box-plot of each 10-fold simple ensemble-average prediction on
accuracy, and MSE for all the 17 models. The red lines are the ensemble-average
prediction with highest accuracy. The blue lines are the other ensemble-average
predictions. The orange lines are the mean accuracy or MSE. The ensemble metric was
either better than or in the upper quantile for all the models. The prediction MSE and
accuracy of each fold are given in table 12 and 13 in appendix C.
232
233
234
235
236
237

Prediction by age class

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
was correctly classified with more than 70% accuracy, and all 13-year-old's was
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).
238
239
240
241
242
243

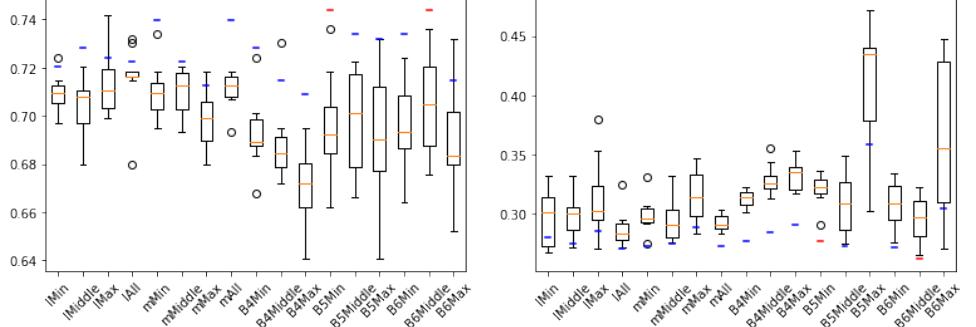


Figure 5. A box-plot of accuracy score (left) and MSE (right) of all the 17 models and the blue line is ensemble-average prediction accuracy (or MSE) on the test-set. The red lines are the two best ensemble-average predictions on accuracy. The orange line are the mean of the 10-fold predictions.

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).

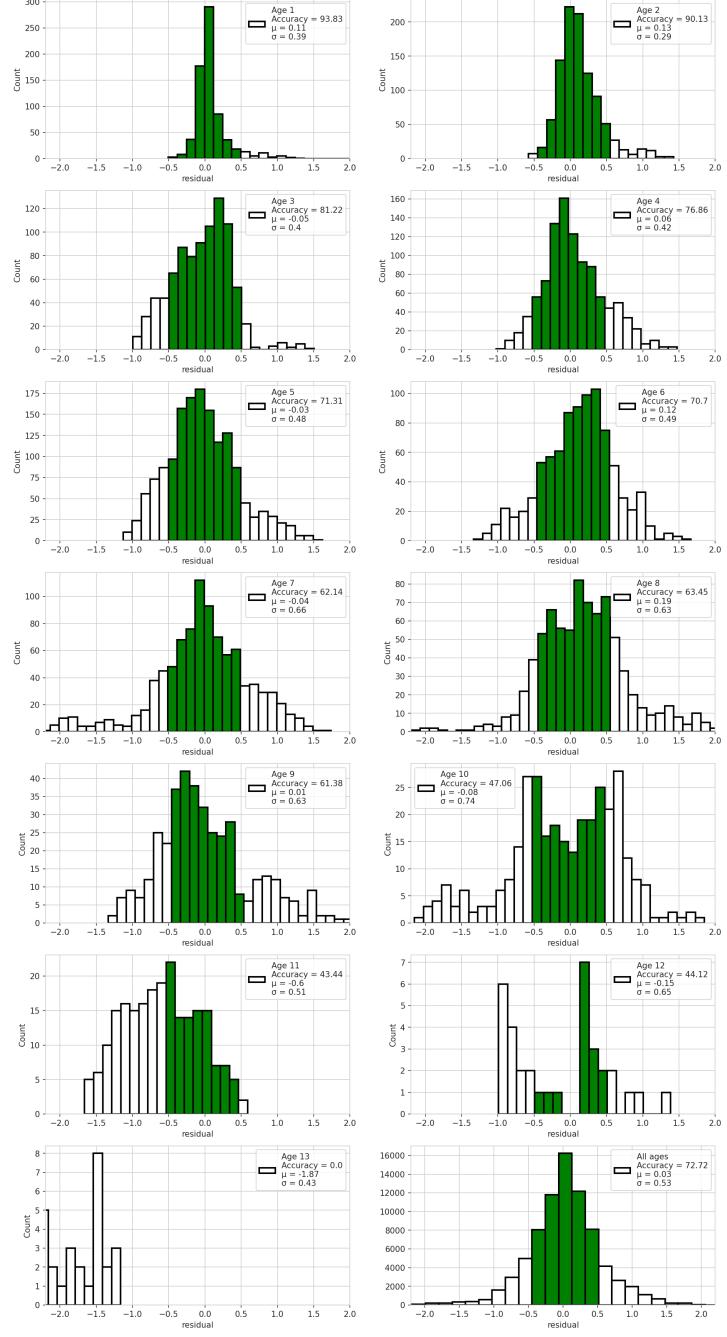
Simple ensemble-average predictions

We searched the space of ensembles-average predictions of 2 to 17 models which is the set of unordered combinations without replacement, equal to the binomial coefficient $\sum_{k=1}^N \binom{N}{k}$ where $N = 17$ and $k \in 2..N$. For each set of ensemble 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.

Table 6 shows the number of combinations which is the number of ensemble models that exists with the coefficient, the best ensemble-average accuracy on the given number of combinations, and the model that produced the best combinations which can be translated to model name using 1. Table 7 shows the same with table but selects the best ensemble-average using MSE.

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

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.



We observe that model B4-min (1), and B6-min (3) was the best models with 14
263 ensembles using them (table 8). These models did not have the highest accuracy of
264 74.6% but an accuracy of 72.8% and 73.4% which was lower than the the highest
265 accuracy models which was B5-min, and B6-middle which had a rank of 3 and 5
266

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

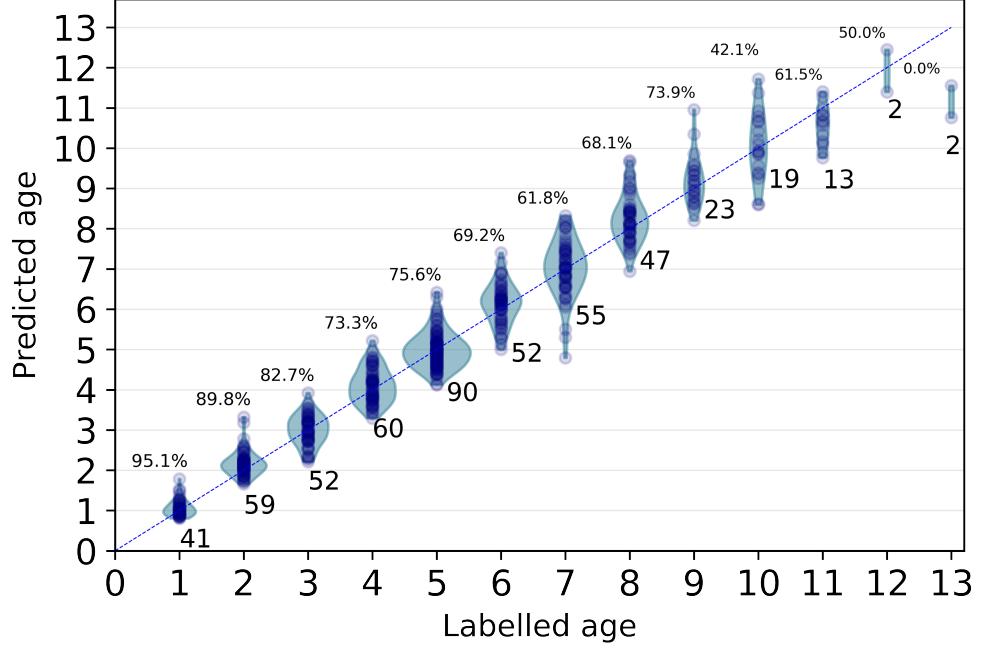


Table 6. Binomial combinations of simple average of ensembles accuracy

Coeff	#Comb	best	model (see table 1)
2	136	75.9	(2, 5)
3	680	77.5	(1, 3, 4)
4	2380	77.9	(1, 2, 3, 4)
5	6188	77.9	(1, 2, 3, 4, 11)
6	12376	78.6	(1, 2, 3, 4, 8, 11)
7	19448	78.1	(1, 2, 3, 4, 7, 8, 11)
8	24310	77.5	(1, 2, 3, 4, 7, 8, 10, 11)
9	24310	77.5	(1, 2, 3, 6, 7, 8, 9, 11, 17)
10	19448	77.1	(1, 2, 3, 6, 7, 8, 9, 10, 12, 13)
11	12376	76.9	(1, 2, 3, 4, 6, 7, 8, 10, 11, 13, 16)
12	6188	76.7	(1, 3, 4, 7, 8, 10, 11, 13, 14, 15, 16, 17)
13	2380	76.3	(1, 3, 4, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17)
14	680	75.9	(1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 16, 17)
15	136	75.7	(1, 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17)
16	17	75.5	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17)
17	1	74.8	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)

respectively.

Exposure types by the number of times in an ensemble were ordered as follows:
min-exposure (rank 4), middle-exposure (rank 9), all-exposures (rank 10), and
max-exposure (rank 11.6).

Table 7. Binomial combinations of simple average of ensembles MSE

Coeff	#comb	best	model (see table 1)
2	136	0.319	(12, 13)
3	680	0.292	(12, 13, 15)
4	2380	0.279	(12, 13, 14, 15)
5	6188	0.272	(12, 13, 14, 15, 17)
6	12376	0.269	(5, 10, 14, 15, 16, 17)
7	19448	0.268	(5, 9, 10, 14, 15, 16, 17)
8	24310	0.267	(4, 5, 9, 10, 14, 15, 16, 17)
9	24310	0.2643	(4, 5, 9, 10, 12, 14, 15, 16, 17)
10	19448	0.262	(4, 5, 9, 10, 12, 13, 14, 15, 16, 17)
11	12376	0.259	(4, 5, 9, 10, 11, 12, 13, 14, 15, 16, 17)
12	6188	0.256	(4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17)
13	2380	0.254	(4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17)
14	680	0.252	(4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)
15	136	0.251	(2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)
16	17	0.250	(1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)
17	1	0.248	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)

Table 8. Rank of number of times a model is in an ensemble

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

271

Table 9 compares the accuracy of the best ensemble with the mean of all 17 models 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 which had 0% accuracy. A distribution plot like figure 5 for the best ensemble can be found in appendix D as figure 16.

272

273

274

275

276

Table 9. 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.

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

278

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 10, with comments by an expert on the most common mispredictions in table 11.

279

280

281

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 was some cod otoliths that was 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).

282

283

284

285

286

287

288

289

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

286

287

288

289

Correlation of predictions and cluster analysis

290

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

291

292

293

294

295

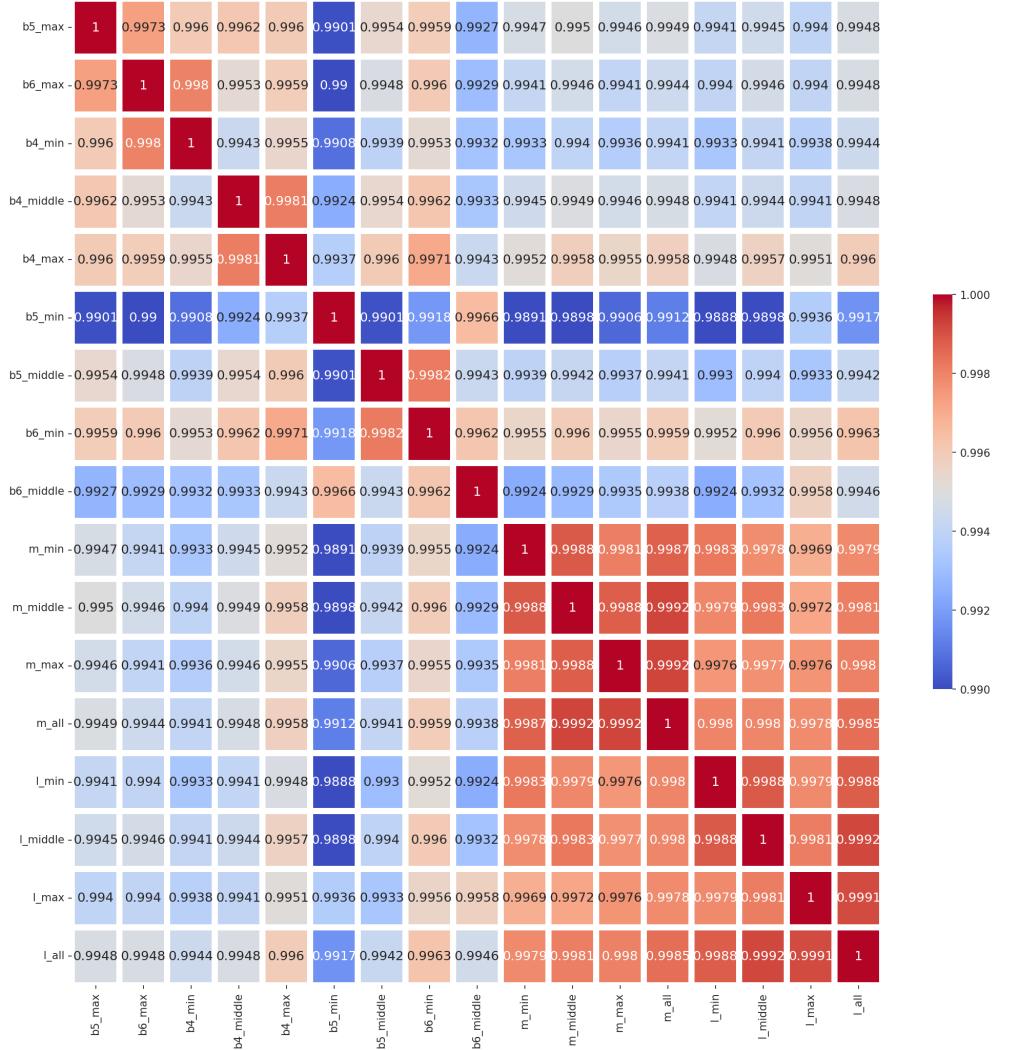
From figure 9 we saw that the EfficientNetV2 family was correlated from the red block in the lower right corner. The dendrogram of figure 10 shows hierarchical

296

297

298

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



clustering (HCA) of the models. HCA found 3 clusters, b5-min, and B6-middle, which
 298 are the two best models, a cluster of all the EfficientNetV2 models, and a cluster of the
 299 rest of the models (figure 10).
 300

Using K-Means together with the elbow- and silhouette-score-methods to find the
 301 optimal number of clusters, they both found that there were 3 clusters in the correlation
 302 matrix (figure 11).
 303

With K-Means and using 3 clusters we found similar clusters to HCA with the
 304 clusters:
 305

- B5-max, and B6-max

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

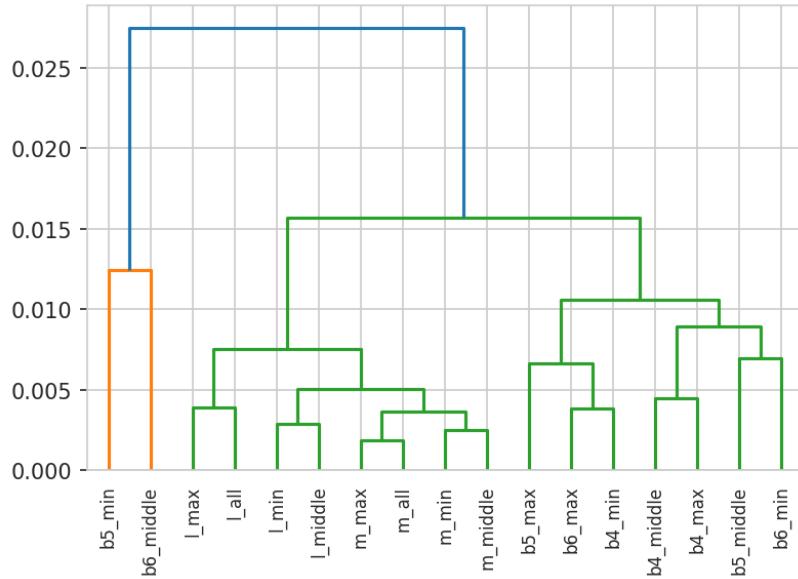
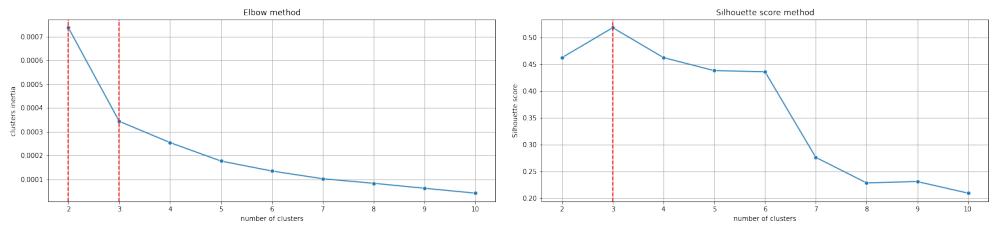


Figure 11. 1.Elbow method and 2. silhouette-score method



- B4-min, B5-min, B6-min, B4-middle, B5-middle, B6-middle, B4-max
- EfficientNetV2 family of models

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

Discussion

During initial training we trained a B4 network on ca 2000 images and obtained an accuracy of ca 60%, later another 3000 images was added and the same network was trained on ca 5000 images which resulted in accuracy of ca 70%. It could be interesting

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

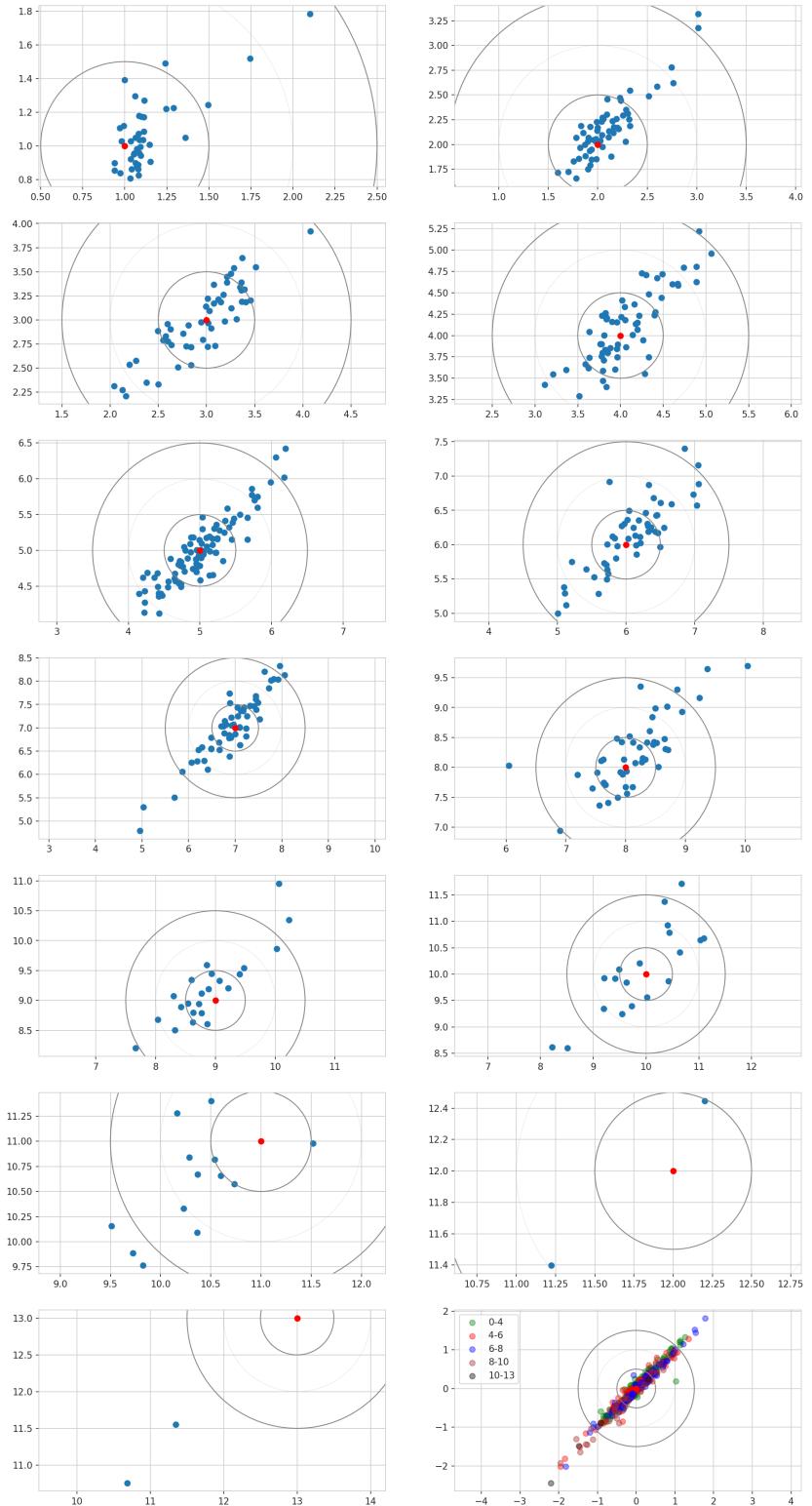
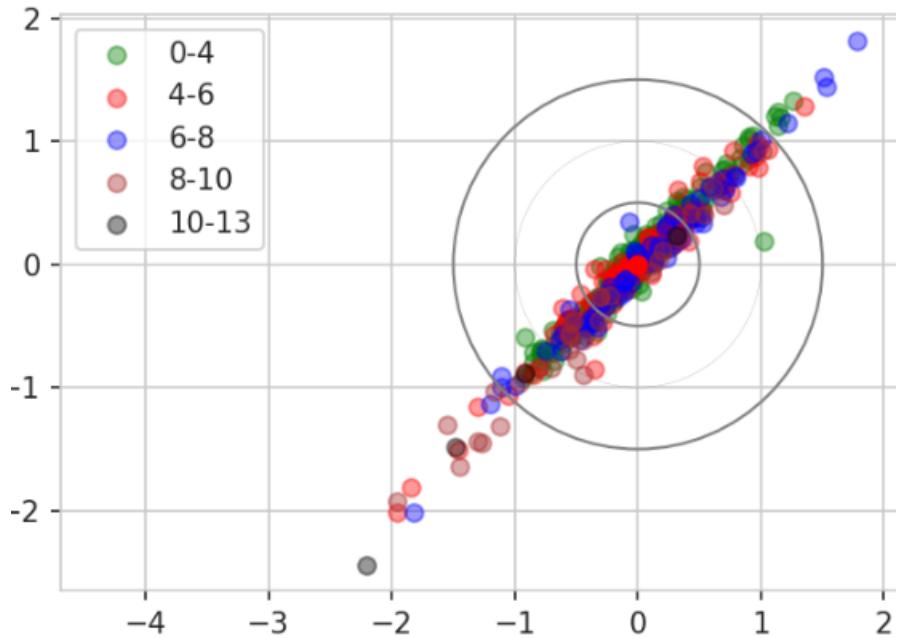


Figure 13. Scatter plot of the residuals of all age classes by Medium-min \times B5-min.



to investigating if adding another 3-5000 images would increase the accuracy to 80%. 318

To reach human level accuracy a score of 85% or higher is required ref-needed, and a 319
score of 90% is considered good. Figure 14 shows a sample of 25 predictions on the 320
validation-set during training. 321

The effect of data size

A crucial issue in machine learning projects is to determine how much training data is 322
needed to achieve a specific performance goal. In computer vision, one commonly used 323

rule of thumb adopted from the number of images and classes in the ImageNet dataset, 324
is to have a thousand images for each class. In our case of cod otolith images the task 325
entails regression towards 13 age classes instead of classification into 1000 classes. 326

Therefore, approximately 13,000 images would appear to be the optimal number for our 327
problem based on the rule of thumb for computer vision. This number can be reduced if 328
transfer learning is applied for images within a similar domain. For cod otolith images, 329
the domain is different than images in ImageNet. However, despite the different image 330
domain for our problem we do see a significant performance boost in using transfer 331
learning, suggesting that fewer images are needed than if trained from scratch. 332

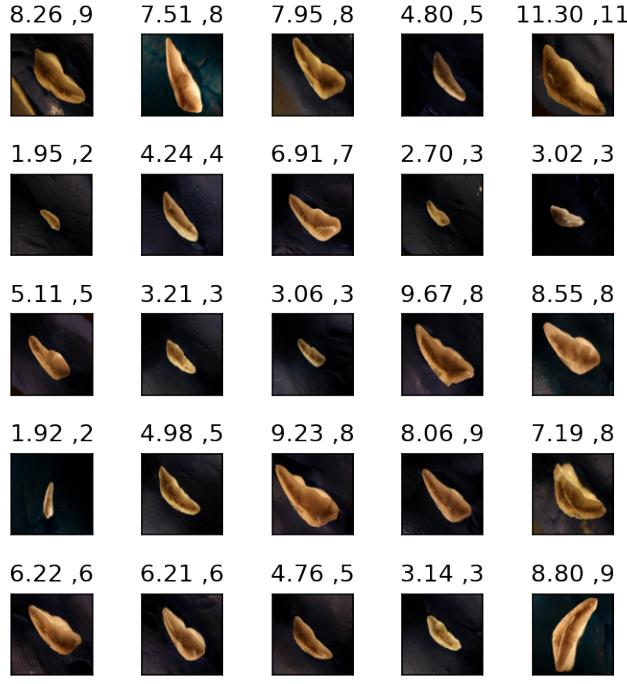


Figure 14. Sample of 25 predictions on a model of training on EfficientNetV2 size medium with minimum light exposure, left number is prediction, and right number is age read

Excessive use of augmentation also reduces the number of images required. On the
 other hand, a general insight from deep learning is that more training data is always
 advantageous. Given the use of transfer learning and augmentation, the number of
 images used in this study, 5000, might be close to the optimal but we still think that a
 larger training set would improve performance. During initial training we trained a B4
 network on about 2000 images and obtained an accuracy of around 60

Accuracy for different age classes

All models tend to predict younger year classes with greater accuracy than older year
 classes. Pooling all models, prediction accuracy for 1–3 year old otoliths is ~80

Ensemble of models

We should discuss the improved accuracy gained when combining models into ensemble
 of models (table 8). The best combinations shows accuracy ~75%. Add a comment on
 the relation between variation in accuracy and the resulting accuracy of the ensemble

(predictions of B4/5/6 have higher variance, but their ensembles win).

347

Moved from introduction to Discussion –

348

Accuracy, efficiency and cost benefits of CNN classifiers versus manual reading 349
... Despite fast progress the results remain mixed and often yield lower precision and 350
consistency than those obtained by trained human readers, which limits the application 351
of automated methods in real conditions. However, one aspect that is often under 352
considered by such studies are the practical time and cost benefits that implementing a 353
functional ML framework would provide. As noted by Fisher and Hunter (2018) (Fisher 354
and Hunter, 2018) in their review of digital techniques for otolith analysis, “costs for 355
human and machine ageing systems are broadly similar since a large part of the cost is 356
associated with preparing the otolith sections”. As such, the net benefit of automated 357
ageing routines is directly dependent on the ability to scale performance using a 358
comparatively smaller number of samples than human readers or, alternatively, to train 359
them on “rougher” data that can be produced faster and at a more efficient cost. Also, 360
CNN can be applied without high additional cost or even be incorporated in the routine 361
protocols, but add a new value e.g., reading consistency check, time-drifts evaluations, 362
inter-reader comparisons (how much ‘off’ is each reader when compared to the CNN 363
predictions, even if not compared with the same otolith samples), etc. 364

365

We see the process of CNN implementation as an evolution of the protocols, with 366
the intensive phase of model development and training. Gradual improvement of model 367
reliability should then allow for the application of CNN as a complementary supportive 368
tool for the age traditional estimations. Finally, this change should aim to scale the 369
capacity of the age reading experts and improve sampling in the areas, fish stocks, or 370
periods that lack proper reading effort.

370

Conclusion

371

Our results demonstrate that the use of deep learning techniques in the analysis of 372
otoliths have a major potential for facilitating automation. We believe that carefully 373
trained CNNs could become a major component in automated pipelines that require 374
minimal processing and could be able to produce near at sea age estimates. 375

When developing the framework for the automatic age estimation it is advised to
376 include B4 architectures as they are quick to train, and performs good. Ensemble
377 approaches are also recommended if more effort is favorable, as it gives a more robust
378 and higher performing prediction. For a quick-to-train ensemble B5, and Medium could
379 be added. It is recommended to use under-exposed images.
380

References

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S.,
383 Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine
384 learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467.*
385
- Albuquerque, C. Q., Lopes, L. C. S., Jaureguizar, A. J., and Condini, M. V. (2019).
386 The visual quality of annual growth increments in fish otoliths increases with latitude.
387 *Fisheries Research*, 220: 105351.
388
- Beamish, R. J. and McFarlane, G. A. (1995). A discussion of the importance of aging
389 errors, and an application to walleye pollock: the world's largest fishery. in recent
390 developments in fish otolith research. *University of South Carolina Press, Columbia,*
391 *S.C.*, pp. 545–565.
392
- Brunel, T. and Piet, G. J. (2013). Is age structure a relevant criterion for the health of
393 fish stocks? *ICES Journal of Marine Science*, 70: 270–283.
394
- Campana, S. (2001). Accuracy, precision and quality control in age determination,
395 including a review of the use and abuse of age validation methods. *Journal of fish
396 biology*, 59(2):197–242.
397
- Carbonara, P. and Follesa, M. C. (2019). Handbook on fish age determination: a
398 mediterranean experience. *General Fisheries Commission for the Mediterranean.
399 Studies and Reviews*, 98: 1–179.
400
- Chollet, F. and others (2018). Keras 2.1.3. <https://github.com/fchollet/keras>.
401

- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE. 402
- 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*. 405
- Fisher, M. and Hunter, E. (2018). Digital imaging techniques in otolith data capture, analysis and interpretation. *Marine Ecology Progress Series*, 598: 213–231. 406
- 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.* 409
- <https://cdnsciencepub.com/doi/abs/10.1139/f04-063> (Accessed 3 February 2022). 412
- 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. 413
- 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. 417
- 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. 420
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. 424
- Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *neurips*. 426
- Mahé, K. (2009). Project no. 044132. *Automated FISH Ageing (AFISA): final activity report.* 428

- Mingxing Tan and, Q. V. L. (2021). Efficientnetv2: Smaller models and faster training. 430
CoRR, abs/2104.00298. 431
- Moen, E., Handegard, N. O., Allken, V., Albert, O. T., Harbitz, A., and Malde, K. 432
(2018). Automatic interpretation of otoliths using deep learning. *Plos One*. 433
- Panfili, J., de Pontual, H., Troadec, H., and Wrigh, P. J. (2002). Manual of fish 434
sclerochronology. <https://archimer.ifremer.fr/doc/00017/12801/> (Accessed 3 435
February 2022). 436
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, 437
Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, 438
M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. 439
(2019). Pytorch: An imperative style, high-performance deep learning library. In 440
Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, 441
R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 442
Curran Associates, Inc. 443
- Politikos, D. V., Petasis, G., Chatzispyrou, A., Mytilineou, C., and Anastasopoulou, A. 444
(2021). Automating fish age estimation combining otolith images and deep learning: 445
The role of multitask learning. *Fisheries Research*, 242: 106033. 446
- Ragonese, S. (2018). Methuselah or butterfly? when fish age estimates and validations 447
tell different stories. the case of the european hake (merluccius merluccius l. 1758) in 448
the mediterranean sea. ””. 449
- Reglero, P. and Mosegaard, H. (2006). Onset of maturity and cohort composition at 450
spawning of baltic sprat sprattus sprattus on the basis of otolith macrostructure 451
analysis. *Journal of Fish Biology*, 68: 1091–1106. 452
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., 453
Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). 454
Imagenet large scale visual recognition challenge. 455
- Siskey, M. R., Wilberg, M. J., Allman, R. J., Barnett, B. K., and Secor, D. H. (2016). 456
Forty years of fishing: changes in age structure and stock mixing in northwestern 457

- atlantic bluefin tuna (*thunnus thynnus*) associated with size-selective and long-term 458
exploitation. *ICES Journal of Marine Science*, 73: 2518–2528. 459
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the 460
inception architecture for computer vision. *CoRR*, abs/1512.00567. 461
- Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional 462
neural networks. *CoRR*, abs/1905.11946. 463
- Tyler, A. V., Beamish, R. J., and McFarlane, G. A. (1989). Implications of age 464
determination errors to yield estimates. *ICES Journal of Marine Science*, 108: 27–35. 465
- Vabø, R., Moen, E., Smoliński, S., Åse Husebø, Handegard, N. O., and Malde, K. 466
(2021). Automatic interpretation of salmon scales using deep learning. *Ecol. Inform.* 467
63, 101322 (2021). 468
- Wightman, R. (2019). Pytorch image models. 469
- <https://github.com/rwightman/pytorch-image-models>. 470
- Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified 471
activations in convolutional network. *CoRR*, abs/1505.00853. 472

A Prediction error of more than 1.5 year

Table 10. 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				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	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	-	

Table 11. Comments on the most frequently miss predicted otolith images

Idx	Comment
13	Labeled 8 years, and read as 10 years by the B-models (EfficentNet). 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 distro 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 EfficentNetV2 models except Medium-middle. The quality of the images was bad. All the exposures was over-exposed.

475

476

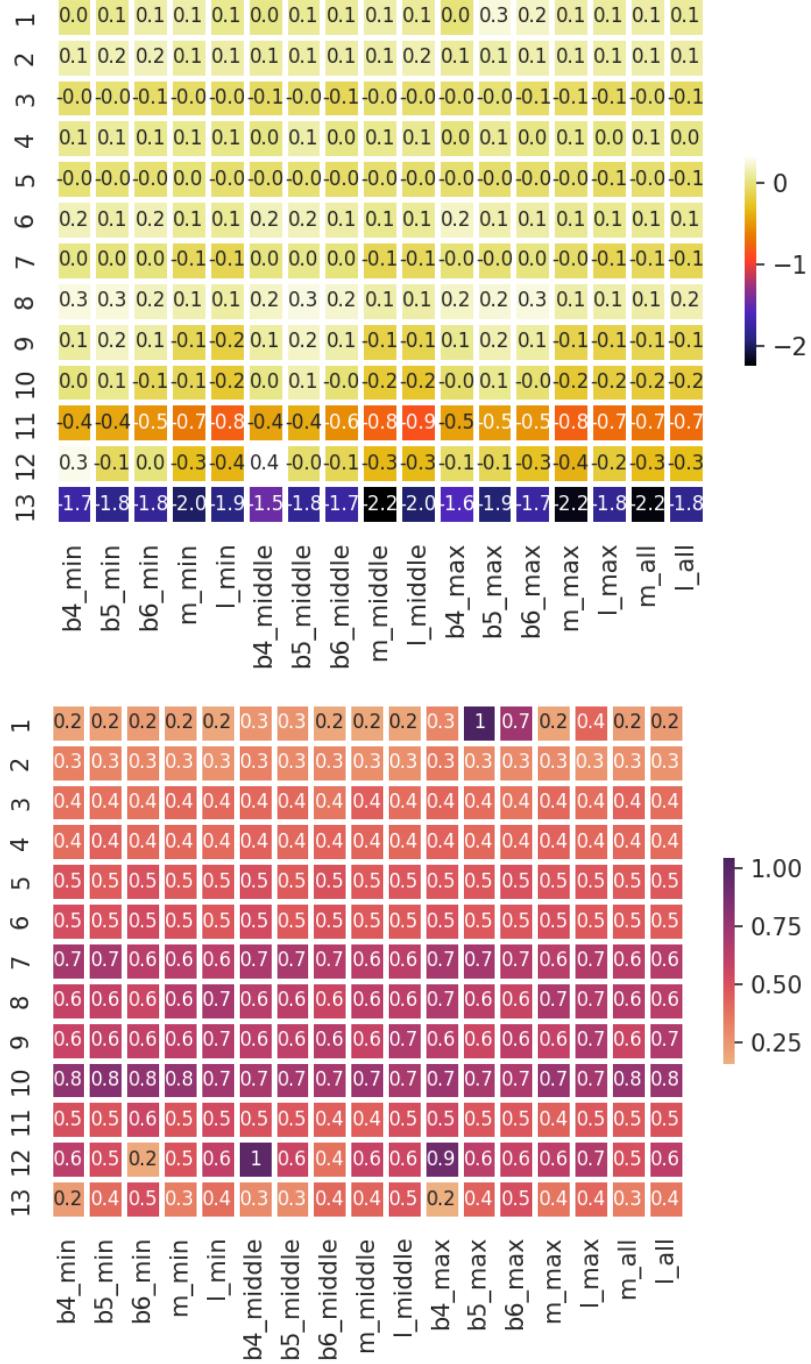
B Model mean and standard deviation of residual

477

test set prediction per age class

478

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



C Model accuracy and MSE per fold

479

Table 12. 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

480

Table 13. 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

482

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

