Automatic interpretation of cod otoliths using deep learning

Endre Moen[1*], Rune Vabø[1], Szymon Smoliński[1], Come Denechaud[1], Ketil Malde[1,2],

1 Institute of Marine Research, Bergen, Norway

2 Department of Informatics, University of Bergen, Norway

* 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, with size distribution, is one of the main criteria used for determining the health of exploited populations and monitoring the effects of selective fishing (4; 14). Changes in the age distribution can track significant changes in population structure, such as a particularly strong year-class skewing the distribution (23), or the gradual truncation of older age classes as selective fishing mortality removes larger individuals (24). 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 (2; 5; 12). While age is inferred from the "simple" counting of annual increments, the interpretation of this zonation pattern is species or even population-specific (15) and is based on precise knowledge of the timing of zone formation and of the correct identification of true and false zones (18). This process therefore requires specific expertise and is subject to uncertainties in both between-reader precision and "true" age accuracy (12). Because those estimates are central to stock assessment, ageing errors or wrong interpretation of

otolith zonation can have dramatic effects on the evaluation of fish biology and consequently stock size and structure (3; 22; 26).

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 (6). This is particularly true for demersal fish species, like Atlantic cod (Gadus morhua), that have large opaque otoliths that can't be read whole and need to be prepared. 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 (17) to the more recent effort toward developing machine learning (ML) frameworks (9; 20). Despite fast progress the results remain mixed and often yield lower precision and consistency than those obtained by trained human readers, which limits the application of automated methods in real conditions. However, one aspect that is often under considered by such studies are the practical time and cost benefits that implementing a functional ML framework would provide. As noted by (11) in their review of digital techniques for otolith analysis, "costs for human and machine ageing systems are broadly similar since a large part of the cost is associated with preparing the otolith sections". As such, the net benefit of automated ageing routines is directly dependent on the ability to scale performance using a comparatively smaller number of samples than human readers or, alternatively, to train them on "rougher" data that can be produced faster and at a more efficient cost.

In this study, we develop a deep learning network for estimating Atlantic cod age using multi-exposure images of broken otoliths set in place using simple plasticine. Our results are positive and show the potential for developing automated pipelines that require minimum processing and could be able to produce near at-sea age estimates.

There are two families of models used, EfficientNet with CNNs B0-B7 (25) and EfficientNetV2 with convolutional neural-networks (CNNs) small, medium, Large, and Xtra-Large (25). The EfficientNet family of models, was introduced in 2019 and the largest model B7 achieved state-of-the-art result on the ImageNet (8) benchmark. It uses neural architecture search to scale image-size and the network. The EfficientNetV2

family of models was introduced in 2021 and Xtra-large achieved state-of-the-art result on the ImageNet benchmark again. It extends on the previous work and introduces new ideas, like scaling up test-set image-size. In this work we investigate EfficientNet B4-B6, and EfficientNetV2 medium and large which shows the best compromise between training-time and accuracy.

Method and materials

Data collection structure should be:

1. Data collection (cruises and archives) and sampling
2. photographic protocol
3. resulting images (size, exposures, number, method)
4. split into datasets and configuration

Data Collection

"1. Data collection and sampling"

We used a dataset sampled from 5150 cod otoliths which has been 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 are sampled using a random-stratified sampling based on fish length for each trawl station, and the otoliths from individual fish are randomly sampled.

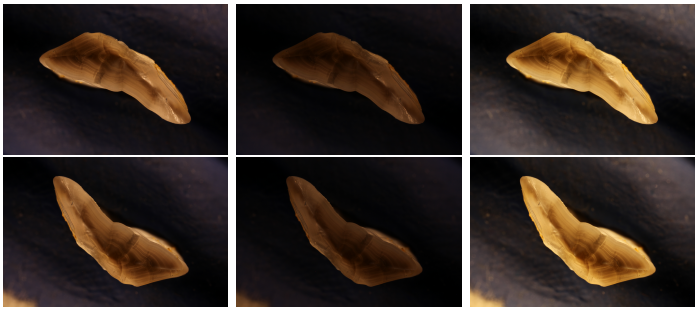
"2. Photographic protocol"

The otolith was broken 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 ventral side of the otolith was near the bottom of the camera.

"3. resulting images (size, exposures, number, method)"

The images are 3744×5616 pixels which are re-scaled for training to between 380×380 and 512×512 pixels. The image light exposure varies depending on light condition outside, and are stored in the metadata of the JPG file. Typically the exposure order is middle-dark-light then the rotation, and then middle-light-dark again.

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



Sometimes the order is changed, so the order is recovered by reading the metadata property.

The details of how the data-set is 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 (10) as well as where the data-set is available.

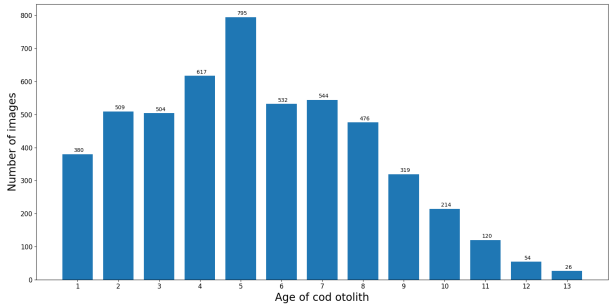


Figure 2. Age distribution of all 5150 images

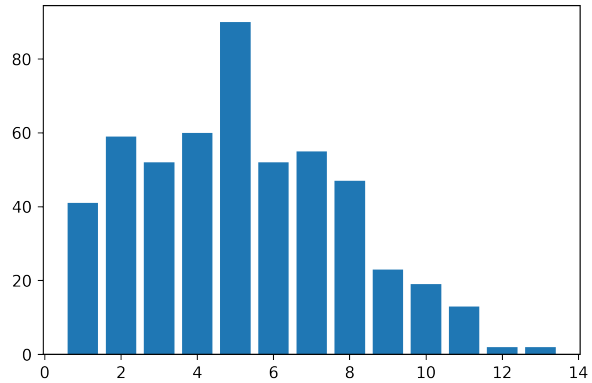


Figure 3. Age distribution of 515 images from the test set

Convolutional neural network architecture

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

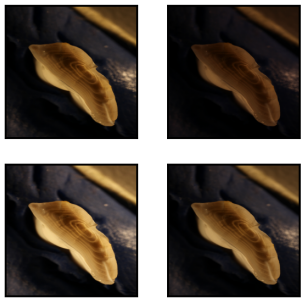
CNN family /	EfficientNet			EfficientNetV2	
Image exposure	B4	B5	B6	Medium	Large
Minimum	1	2	3	4	5
Medium	6	7	8	9	10
Maximum	11	12	13	14	15
All (3 images)	-	-	-	16	17

Each CNN was trained using transfer learning by loading ImageNet weights. The image size varies between 380×380 and 528×528 pixels. While test-set size prediction has been done on 380×380 and 384×384 pixels. To investigate the image-taking protocol described in (10) we have also training on 9-channel images. Three RGB-images are stacked to produce a 9-channel image. Using Timm(27) the imageNet weights were duplicated on the input layer to accommodate 9 channels. The 3 images used are of dark, medium and light exposure of the first orientation.

CNNs was selected based on performance on the ImageNet benchmark and availability of open-source implementations with imageNet weights. The imageNet benchmark is for classification while we treat aging as a regression problem (9) (R. et al.). The last layer of the CNNs has been modified to output a linear output. In the EfficientNetV2 family we have done this by applying three multi-layer perceptron layers going from 1280 output of last hidden layer to dense 256-layer, then a leakyRelu (28) layer, and then dense 32-layer, then a leakyRelu layer, and finally a linear output layer. For EfficientNet we only change the last layer from softmax output to a linear output.

To each fold we normalize the age on the training-set by removing the mean and scaling to unit variance. The normalization is then applied to validation and test-set using sklearns StandardScalar. Test-set predictions are obtained by applying the inverse transform.

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).



Implementation and training

EfficientNetV1 B4, B5, and B6 was implemented with TensorFlow (1) and Keras (7) software packages in Python. Computation was done using CUDA 11.1 and CuDNN with Nvidia(Nvidia Corp., Santa Clara, California) A6000 accelerator card with 48 GB of GPU memory and P100 cards with 12 GB 112 of GPU memory, EfficientNetV2 medium, and large was implemented with the PyTorch (19) and timm (27) software package. Computation was done on P100 and RTX 3090 with 24 GB of GPU memory. Pretrained weights for EfficientNet was available from Keras, and pretrained weights for EfficeintNetV2 was available from Timm.

Augmentation was applied to the training-set. The images were augmented using rotation between 0 and 360 degrees, and reflection by the vertical axis. The pixel values has a range between 0 and 255 which was normalized to between 0 and 1.

The augmented data set can produce $360 \times 2 \times 5150 = 3.708.000$ possible images. Depending on the augmentation factor and the number of images in a training cycle, the model will likely never see the same image twice.

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

To get the most out of a small data-set we applied 10-fold cross-validation on 90% of the data-set, 4635 otoliths. Each fold of the 10 folds consists of 90% of the cross-validation set and 81% of the whole data-set, 4172 otoliths for training. Each fold had then 463 otoliths for validation which is 10% of the cross-validation set, and 9% of

the whole data-set. Each model is training on the 4172 otoliths and the model with the best MSE on the 463 otoliths in the validation set is chosen. The best model on the validation set was then used to predict the age on the test-set, and the metric for accuracy and MSE was recorded. The test-set is chosen at random, while the 10-fold split is chosen using stratified-kfold split which preserves a similar distribution of the whole cross-validation set in each validation set. That means the 463 images in the validation-set will have similar age distribution to that of the 4635 images in the cross-validation set. Both the test-set and the whole data-set follows a normal distribution with largest age-class being 5-year-olds.

Hyper-parameters

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

EfficientNet uses learning-rate with no scheduler while EfficientNetV2 uses Cosine Annealing scheduler (16). The training- and validation image size is as described in the papers except for Large which uses 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).

Table 2. Hyper-parameters on each model

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

Medium all-, and min-exposures was run with `steps_per_epoch`=160
B6 has epochs=150, `early_stopping_patience`=14, and `reduceLROnPlateau_patience`=7
B4 min was run with `img_size`=456

Table 3. Hyper-parameters on all models

learning_rate	1e-05
n_fold	10
test_size	0.1
in_chans	3 or 9

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

Table 4. Hyper-parameters on TensorFlow models (B4, B5, and B6)

reduceLROnPlateau_factor	0.2
which_exposure	min, medium, max

Table 5. Hyper-parameters on PyTorch models (medium, large)

scheduler	CosineAnnealingLR
T_max	10
min_lr	1e-06
weight_decay	1e-06
which_exposure	min, medium, max, all

We trained 10 models using 10-fold cross-validation which produced an ensemble prediction based on the test-set prediction on the test-set. Typically the ensemble prediction is better than any single fold prediction. Ensembles are better because they improve performance. An ensemble can make better predictions and achieve better performance than any single contributing model, just as more experts will produce higher accuracy in predicting a single otolith. Robustness; An ensemble reduces the spread or dispersion of the predictions and model performance. This result can be improved further by taking ensemble predictions of ensembles. We look at all ensembles from tuple-ensembles, consisting of 2 models, which produces an ensemble of 20 models, and triplet-ensembles consisting of 3 models, to ensemble of all models which produces an ensemble consisting of 170 models.

By choosing the best model we are over fitting to the test-set, but selecting a subset of the best of these ensembles should produce a candidate ensemble of ensembles which will likely produce the best prediction on a hold-out test-set.

Simple ensemble learning with averaging

Ensemble averaging is a simple form of committee machines (13). We investigate both simple mean and weighted mean of the 10-fold ensemble models. Simple mean gives each model equal importance and weighted mean is represented by a set of weights that sum to 1.0.

Why does ensembles work? Assume we measure a random variable (x), with a normal distribution, which is denoted as $\mathcal{N}(\mu, \sigma^2)$ with μ, σ the mean and standard deviation.

Measuring only one variable once, we know $\mathbb{E}[x_1] = \mu$ and $Var(x_1) = \sigma^2$ for any $x_1 \in (x)$

Suppose we measure the random variable (x), P times $(x_1, x_2..., x_p)$. That is, measurement in the form of $(x_1, x_2..., x_p)/P$. Then the mean will still be μ. However, the variance will be smaller:

$$Var(\frac{x_1 + ... + x_p}{P}) = \frac{Var(x_1) + ... + Var(x_p)}{P^2} = \frac{P\sigma^2}{P^2} = \frac{\sigma^2}{P}$$

So the mean stays the same, while the variance is averaged. Hence the variance is reduced.

Clustering analysis

How much does the CNNs agree with each other and which CNNs are more in agreement? The first question can be answered by looking at the correlation matrix on the predictions on the test-set and the second question can be answered by a cluster analysis on the correlation matrix.

We will use Pearson's correlation coefficient as the relations between the predictions on the test-set is linear, and for clustering we will look at hierarchical clustering (HCA) and K-means clustering with the number of clusters given by the elbow-, silhouette-score-method. In HCA we use euclidian distance (Chebyshev, and Minkowski

gave same results), and we used Complete-linkage clustering.

Results

In table 1 and table 2 are the accuracy and MSE metrics for ensemble predictions on the 10-fold training. It can be observed that in the EfficientNet family, larger networks has better MSE, while accuracy is not as correlated. A similar pattern can be observed for the EfficientNetV2 networks. However it seems like EfficientNet is better than EfficientNetV2 in both metrics unlike the results observed on the ImageNet benchmark. The two best models are both from EfficientNet, B5-min and B6-middle exposure score 74.4% accuracy on the test-set.

Table 6. Accuracy by light exposure and CNN architectures

MSE:light/CNN	B4	B5	B6	Medium	Large
min	72.8*	74.4**	73.4*	74.0	72.0
medium	71.5	73.4	74.4**	72.4	72.8
max	70.9	73.2	71.5	71.1	72.4
9 channels	-	-	-	74.0	72.2

* should be retrained
** best result

Table 7. MSE by light exposure and CNN architectures

ACC:light/CNN	B4	B5	B6	Medium	Large
min	.277	.277	.272	.273	.280
medium	.285	.273	.262	.292	.275
max	.291	.359	.305	.290	.286
9 channels	-	-	-	.273	.271

We compare the 10-fold prediction accuracy, and MSE of all the models in a box plot in figure 5. The red line is the ensemble score with the highest accuracy. The blue lines are the other ensemble accuracies or MSE's. The orange line is the mean accuracy or MSE. The ensemble metric is either better than or in the upper quantile for all the folds, and always better than the mean prediction (orange line).

By comparing the models on MSE we can see that larger models are better, e.g B6 has higher mean than B5 and B4, and large is better than medium. We also see that

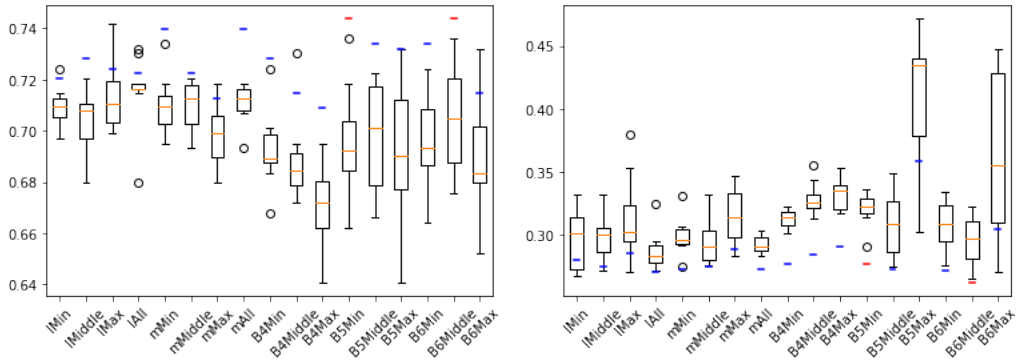


Figure 5. Accuracy score(left) and MSE of all the 17 models and the red line is simple ensemble-average prediction accuracy

the EfficientNetV2 networks has higher mean than the first generation EfficientNet. However, this is not true for the ensemble predictions (red line) nor for the fold-mean or ensemble of the accuracy. We can also see that the effect of adding 3 images, creating 9 channels, on the model is that the variance is reduced, the fold mean metric increases, but the ensemble metric is reduced.

The box plots are produced from the folds given in table 12 and 13 in appendix C

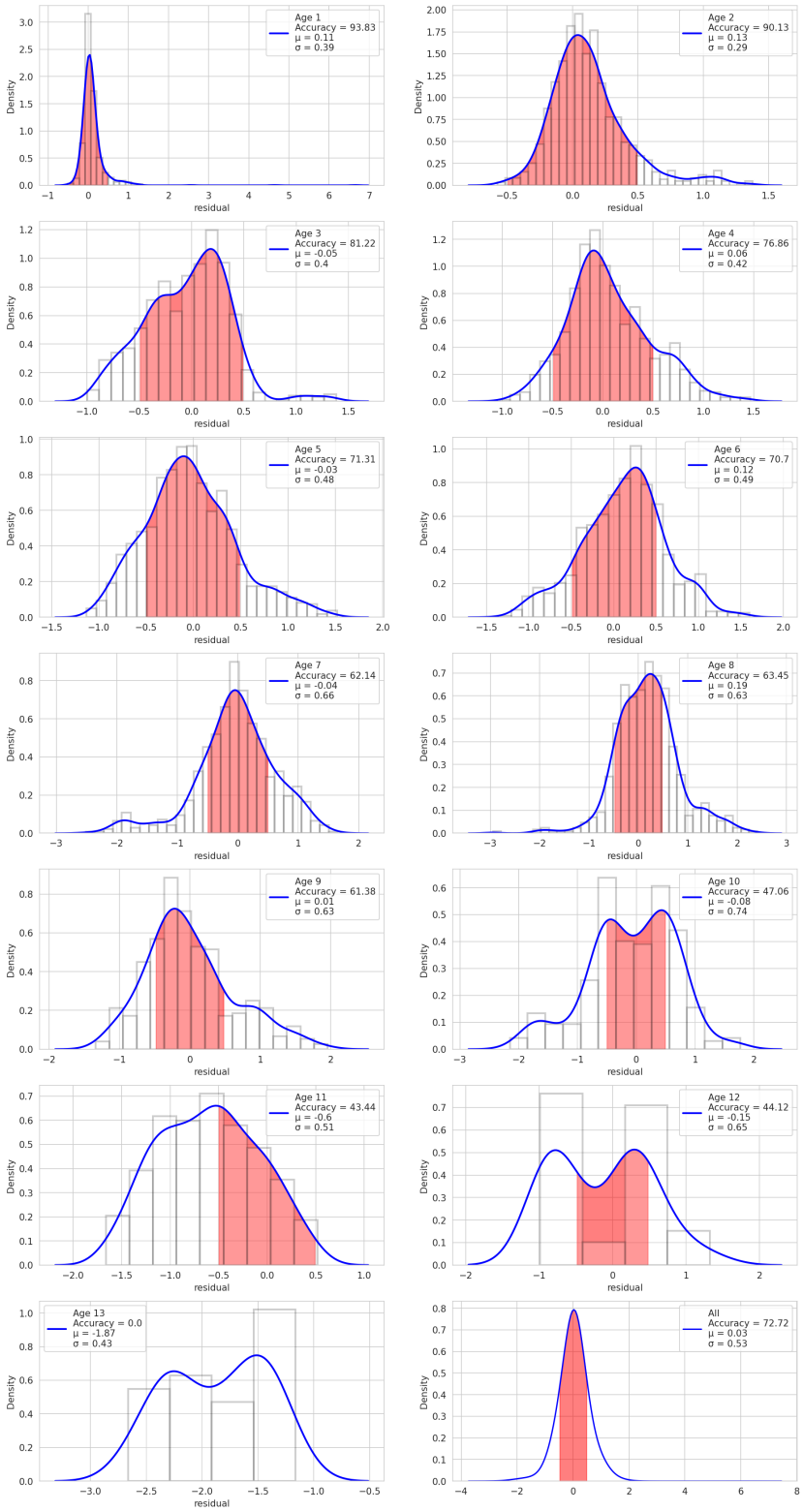
Prediction by age class and residuals

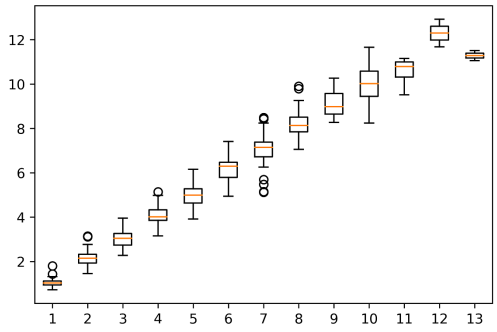
Figure 6 shows the residual error as the average across all models. It looks like the residuals follow a normal-like distribution. Assuming all age groups for all models are normal distributed, a table with mean and standard deviation can be found in Appendix B for each model and age-group

The figures below 7 shows the predictions per age group on the test-set. We can see that the prediction follows a linear trend $y = x$ except for the 2-3 last years, when the mean drops below $y = x$. This is even more obvious in the residual plots where the prediction drops below $y = 0$ for the 12- and 13 year-old's. The models has a bias towards lower age for these age groups which is a sign of under-fitting. This correlates with the limited number of otolihts in the oldest age groups, and could be explained by reversion to the mean, as the largest age-group in is 5-year-olds.

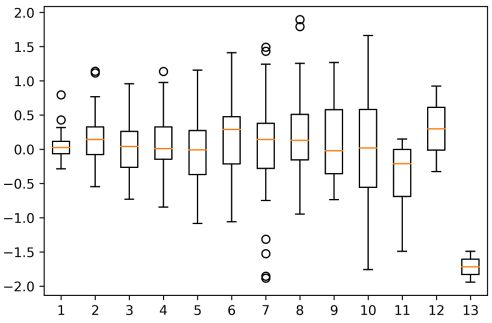
Figure 8 shows scatter plots of all predictions that results in a miss-classification. That is predictions that has an error greater than 0.5 in magnitude. Predictions that miss by more than 1.5 in magnitude are shown with red dots.

Figure 6. Residuals per age class over all models. Red region is correctly classified

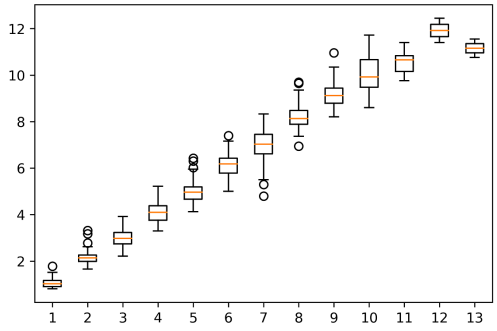




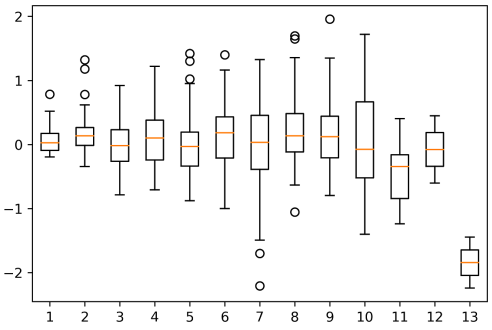
(a) B4 minimum exposure



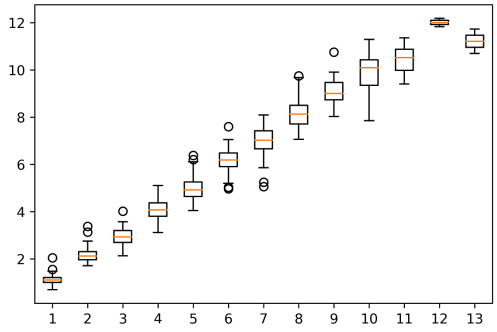
(b) B4 minimum exposure residuals



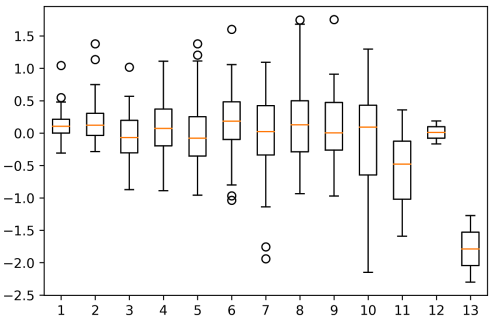
(c) B5 minimum exposure



(d) B5 minimum exposure, residuals

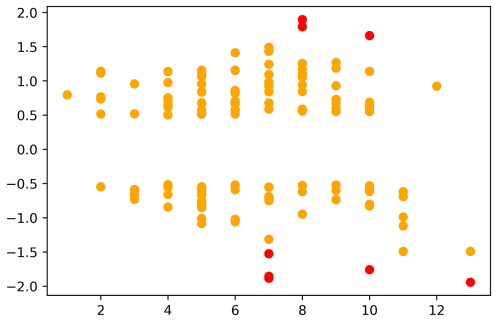


(e) B6 minimum exposure

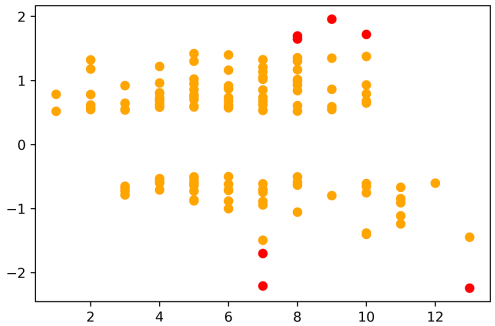


(f) B6 minimum exposure, residuals

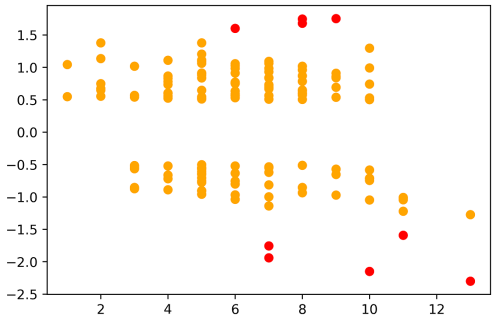
Figure 7. Shows the erroneous predictions per age group on the test-set, for each model



(a) B4 minimum exposure



(b) B5 minimum exposure



(c) B6 minimum exposure

Figure 8. Comparing the models, looking at age per age class, and the reciduals per perdiction

Ensembles with averaging

We search the space of ensembles with simple-average and weighted-average predictions which are given by $\sum_{k=1}^{N} \binom{N}{k}$ where $N = 17$ and $k \in 1..N$ and find that the best ensemble is given by six models and produce an accuracy of 78.6%. The model consists of models (1, 2, 3, 4, 8, 11) which is (using table 1) B4-min, B5-min, B6-min, Medium-min, B6-middle, and B4-max. The accuracy goes down after adding 6 models, while the MSE continue to improve until all 17 models are included. This is as expected from the theory of simple ensemble learning, which states that the more use measure a random variable the lower the variance is.

Table 8. Binomial combinations of simple average of ensembles accuracy

comb.	`#comb`	best	model
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)

If we rank how many times a model is included in an ensemble (See table 10):

then we see that model 1, and 3 are the best with 16 ensembles including them. Model 1 and 3 is model B4-min and B6-min, but these models has an accuracy of 72.8% and 73.4% which is lower than the highest accuracy of 74.6% produced by model B5-min, and B6-middle which has a rank of 14 and 13 respectively.

Table 9. Binomial combinations of simple average of ensembles MSE

comb.	#comb	best	model
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 10. Rank of number of times model is in an ensemble

Model no.	Model name	Count
1	B4-min	16
2	B5-min	14
3	B6-min	16
4	Medium-min	14
5	Large-min	8
6	B4-middle	5
7	B5-middle	11
8	B6-middle	13
9	Medium-middle	7
10	Large-middle	9
11	B4-max	11
12	B5-max	5
13	B6-max	8
14	Medium-max	6
15	Large-max	5
16	Medium-all	7
17	Large-all	7

Outliers

Looking at figure 7 we can see that the model under-predicts the age of older otoliths. This pattern is especially observable for individuals read as 13 years. To better understand the bias, figure 9 shows 6 images which has an error of more than 1 year. The index of the images in the test-set is (13, 71, 270, 342, 360 and 369). Which networks made the miss-prediction and by how much as well as other images that had a prediction error of more than 1 year can be found in table 11 in appendix A. From the

table we can make the following observations:

- some images are outliers to all models (71, 342, 362, 369)
- some images are outliers to families of models (13, 423)
- some images are outliers based on light exposures (308)
- some images are just randomly outliers (320)
- the number of large outliers doesn't look like it correlates with model performance (B5-min, B6-mid)
- Image 308 is a large outlier for B6 max, B5 max, and large max.

Figure 9. Some of the most common images with miss-predicted of more than 1.5 years

Correlation of predictions and cluster analysis

From the outliers we can see there is a correlation of predicting outliers across models. Lets look at the correlation of models on the test-set predictions.

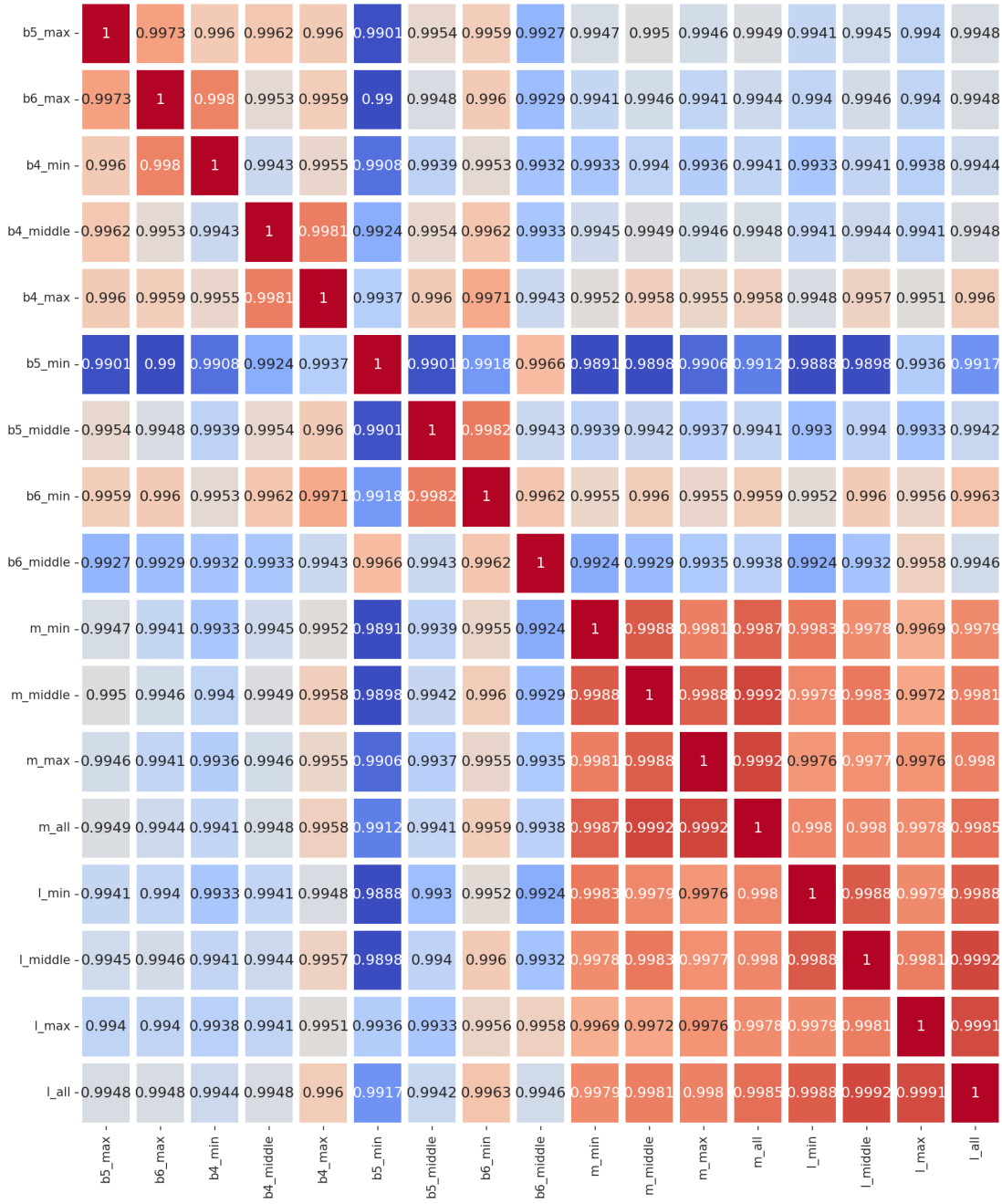
We can see that EfficientNetV2 models are most correlated to each other, and that B4 models are correlated. Also B6-middle is correlated to all models. All the results are highly correlations.

Lets look at HCA:

And lets see how many clusters there should be with the elbow- and silhouette-score-methods.

These methods suggests that there should be 3 clusters. Using K-Means we find that these clusters should be: (B5-max, B6-max), (B4-min, B5-min, B6-min, B4-middle, B5-middle, B6-middle, B4-max), and all exposures of (Medium, Large).

Figure 10. Pearson correlation of each model prediction on the test-set



We can also look at the correlation between models pr age-class. 7 shows scatter plots one of the least correlated models, Medium-min and B5-min which is one of the least correlated CNNs with a pearson's correlation of 0.988. The last figure is the residual correlation of all age-groups.

Figure 11. HCA

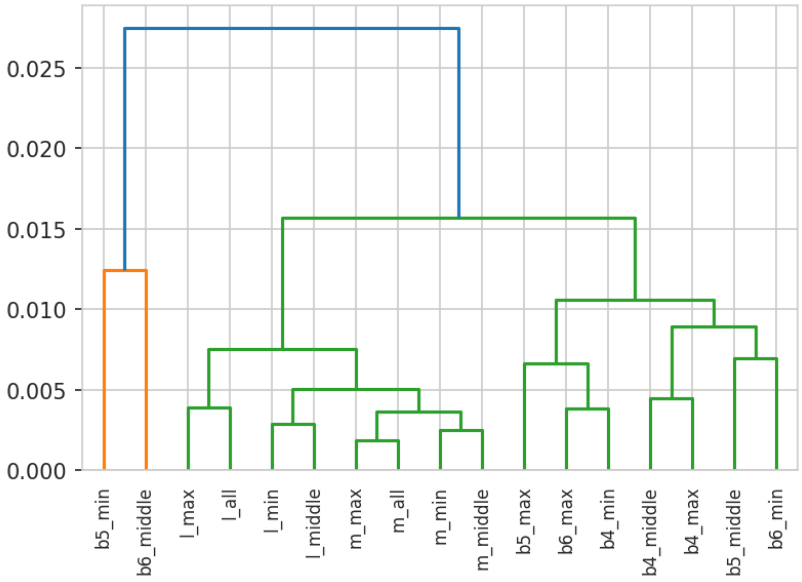


Figure 12. 1.Elbow method and 2. silhouette-score method

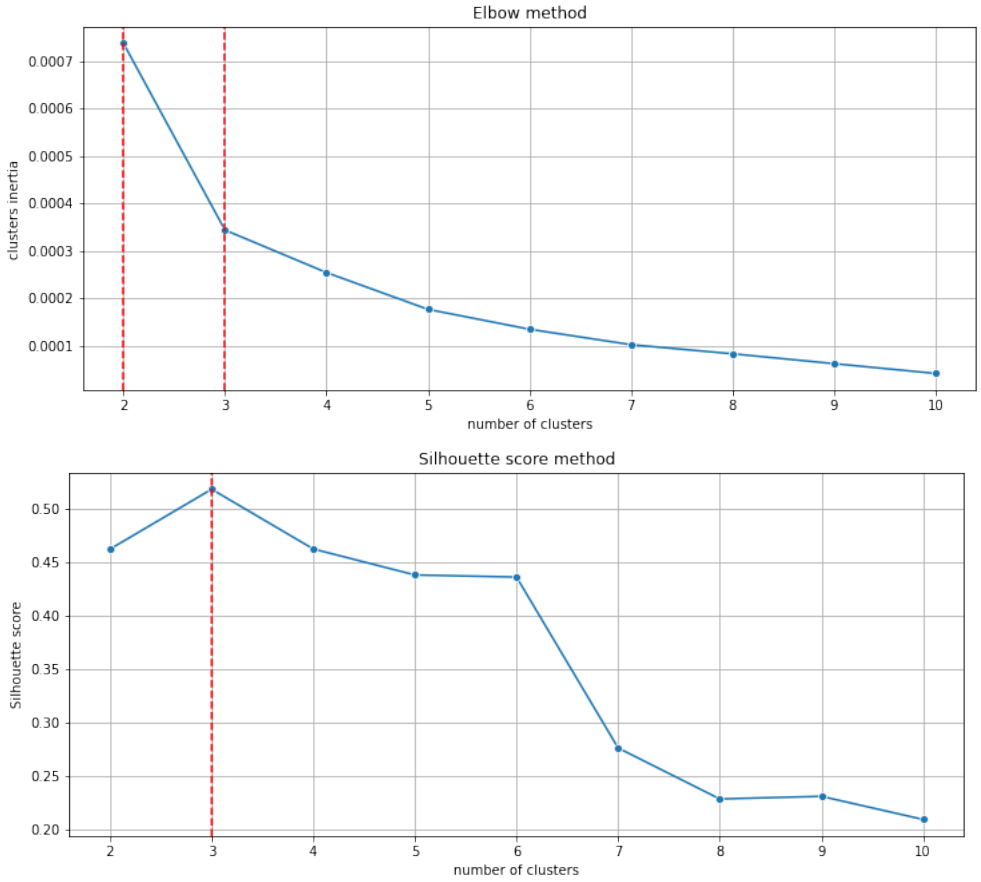
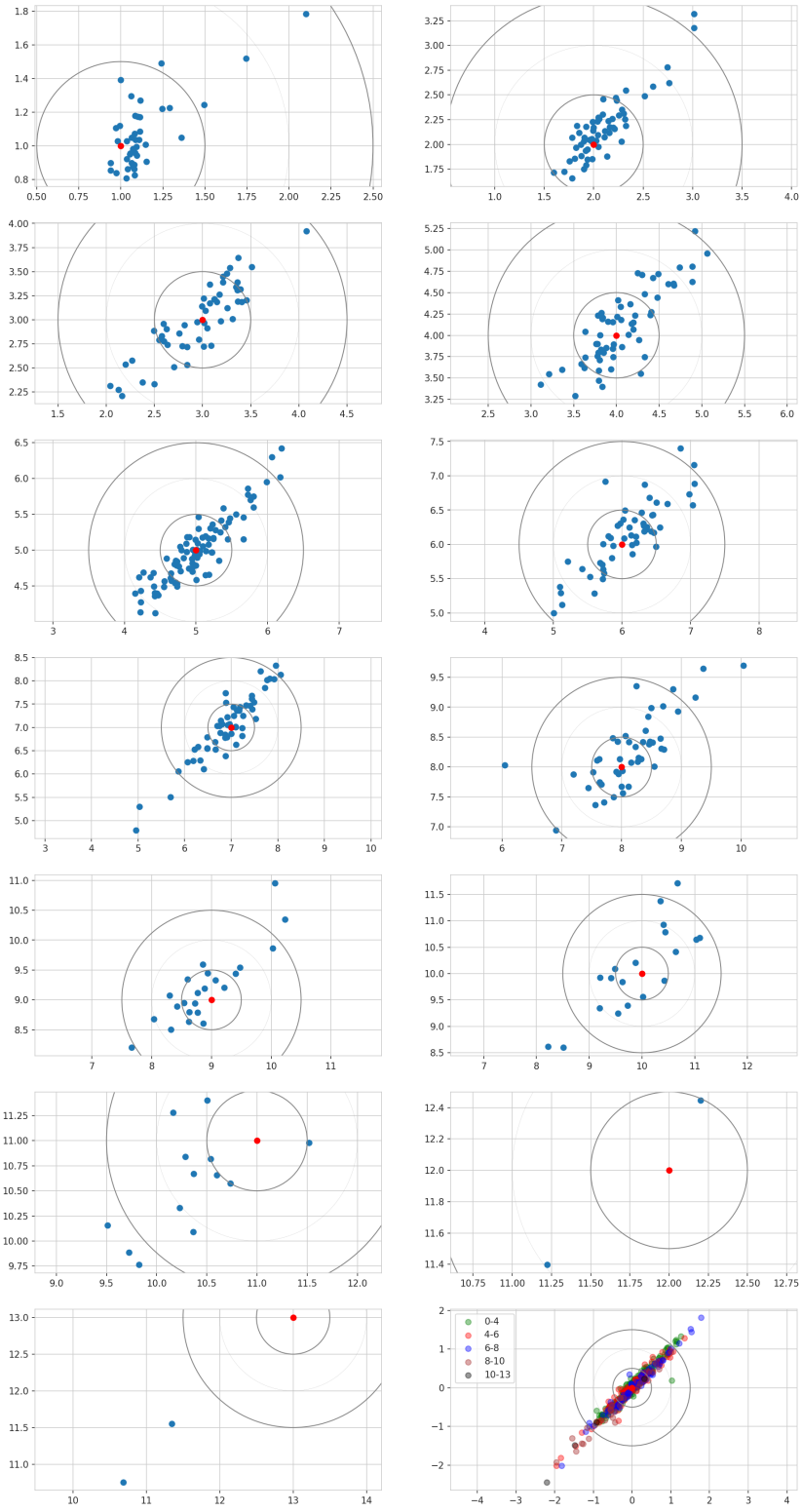


Figure 13. Scatter plot of each age-class by Medium-min × B5-min.



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 to investigating 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 (**?**), and a score of 90% is considered good.

Why is image with index 308 in the test-set such a large outlier for the largest models B5, B6 and Large with max exposure?

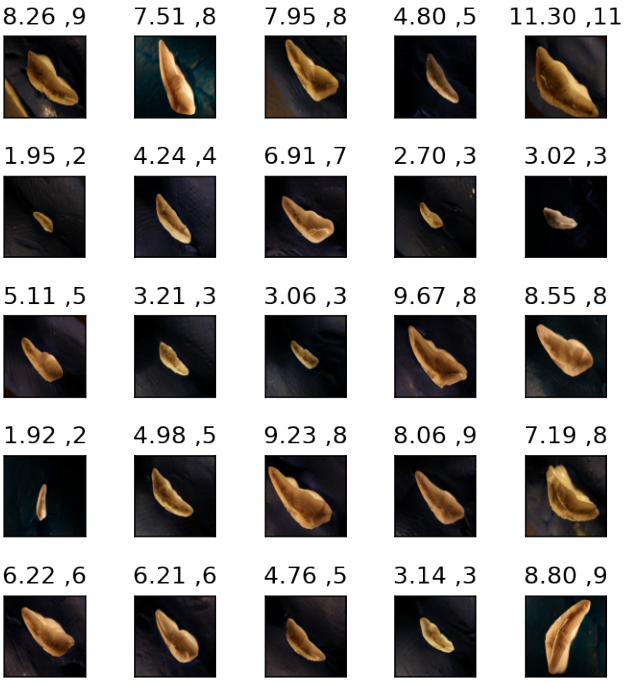


Figure 14. Sample of 25 predictions on a model of training on EfficientNetV2 size medium with minimum light exposure, left number is prediction, and right number is age read

References

References

1. 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*.

2. 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*.

3. 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.*

4. 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*.

5. 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.

6. 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.*

7. Chollet, F. and others (2018). Keras 2.1.3. https://github.com/fchollet/keras.

8. 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.

9. E., M., Handegard, N. O., Allken, V., Albert, O. T., Harbitz, A., and Malde, K. (2018). Automatic interpretation of otoliths using deep learning.

10. 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*.

11. Fisher, M. and Hunter, E. (2018). Digital imaging techniques in otolith data capture, analysis and interpretation. *Marine Ecology Progress Series, 598: 213–231.*

12. 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).*

13. HAYKIN, S. (1999). Neural networks - a comprehensive foundation. *Second edition. Pearson Prentice Hall*.

14. 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.*

15. 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.*

16. Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts.

17. Mahé, K. (2009). Project no. 044132. *Automated FISh Ageing (AFISA): final activity report.*

18. 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).*

19. 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.

20. 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.*

R. et al.. R., V., Moen, E., Smoliński, S., Åse Husebø, Handegard, N. O., and Malde, K. Automatic interpretation of salmon scales using deep learning. *Ecol. Inform. 63, 101322 (2021).*

22. 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. *"".*

23. 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.*

24. 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.*

25. Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946.

26. 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.*

27. Wightman, R. (2019). Pytorch image models. `https://github.com/rwightman/pytorch-image-models`.

28. Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853.

A Common outliers of more than 1.5 years, 'm' is Medium, and 'l' Large network

*: B5 max has a large outlier on image 308 with read age 1 years, and predicted age 7.7 years.

**: B6 max has a large outlier on image 308 with read age 1 years, and predicted age 7.7 years.

1. some images are outliers to all models (71, 342, 362, 369)
2. some images are outliers to families of networks (13, 423)
3. some images are outliers based on light exposures (308)
4. some images are just randomly outliers (320)
5. number of large outliers doesnt look like it correlates with model performance (B5-min, B6-mid)
6. Image 308 is a large outlier for B6 max, B5 max, and large max

B Mean and standard deviation per model x per Age group

C Accuracy and MSE per model and per fold

Table 11. Outliers with more than 1.5 year error. Prediction and true age, per model

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	9
B6-mid	9.6				5.1			11.5	9.7					10.8	5.2	8.3	10.8						9.4	9
B6-max	9.8				5.2					5.7				10.7	5.2	8.2	10.6				6.5		9.4	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			5.1				9.9	3.6		5.4		10.8	5.1	8.2			5.9			8.4		10
l-all		9.3			5.0				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	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	100	35	12	41	12	6	18	18	0

Table 12. MSE per CNN and per fold, 'm' is Medium, and 'l' Large network

CNN/fold	1	2	3	4	5	6	7	8	9	10	ens.
B4,min	.320	.318	.306	.313	.322	.314	.315	.316	.306	.302	.277
B4,middle	.344	.328	.316	.334	.326	.320	.355	.326	.313	.325	.285
B4,max	.340	.317	.318	.347	.336	.336	.336	.320	.354	.336	.291
B5,min	.324	.322	.325	.336	.291	.314	.320	.331	.33	.317	.277
B5,middle											
B5,max											
B6,min	.325	.329	.334	.293	.312	.290	.320	.300	.276	.306	.272
B6,middle	.323	.301	.312	.268	.294	.266	.309	.311	.278	.289	.262
B6,max	.435	.306	.306	.270	.390	.321	.411	.321	.294	.448	.305
m,min	.292	.292	.294	.275	.298	.304	.304	.331	.307	.295	.273
m,middle	.321	.377	.332	.285	.285	.325	.311	.348	.295	.373	.292
m,max	.337	.297	.302	.291	.315	.347	.338	.321	.313	.283	.289
m,all	.292	.289	.289	.326	.307	.327	.283	.300	.335	.295	.281
l,min											
l,middle	.301	.281	.299	.318	.282	.305	.280	.334	.3	.310	.280
l,max											
l,all	.292	.289	.289	.326	.307	.327	.283	.30	.335	.295	.281

Table 13. Accuracy per CNN and per fold, 'm' is Medium, and 'l' Large network

CNN/fold	1	2	3	4	5	6	7	8	9	10	ens.
B4, min	69.9	68.9	68.7	68.3	68.9	70.1	69.7	66.8	68.9	72.4	72.8
B4, middle	68.5	69.3	73.0	68.5	67.8	68.2	67.2	67.2	68.3	69.5	71.5
B4, max	64.1	68.2	67.2	66.2	67.8	69.5	67.2	69.3	66.2	65.2	70.9
B5, min	71.8	69.1	69.3	66.8	73.6	70.7	66.2	68.3	69.5	68.7	74.4
B5, middle											
B5, max											
B6, min	68.3	68.5	66.4	72.4	70.7	70.9	69.3	69.3	72.0	68.9	73.4
B6, middle	68.5	69.9	67.6	73.6	72.8	72	68	69.3	72	71.1	74.4
B6, max	70.5	68.2	65.2	73.2	69.1	67.8	68.0	68.0	72.8	68.5	71.5
m, min	71.1	71.1	69.5	73.4	71.8	70.9	70.9	69.7	70.1	71.5	74.0
m, middle	68.7	67.6	68.3	71.1	70.1	70.5	69.9	68.3	69.9	66	72.4
m, max	68.9	70.1	70.3	71.3	70.7	68.5	69.7	68.0	69.1	71.8	71.3
m, all	71.7	70.7	69.3	71.3	71.8	71.8	71.3	71.7	71.1	70.7	74.0
l, min											
l, middle	69.7	73.4	69.1	67	71.8	69.9	72.6	68.2	70.5	70.3	71.8
l, max											
l, all	70.9	70.7	70.5	70.7	71.5	69.3	70.7	71.8	69.7	70.9	71.7

Figure 15. Mean of residuals per age-group

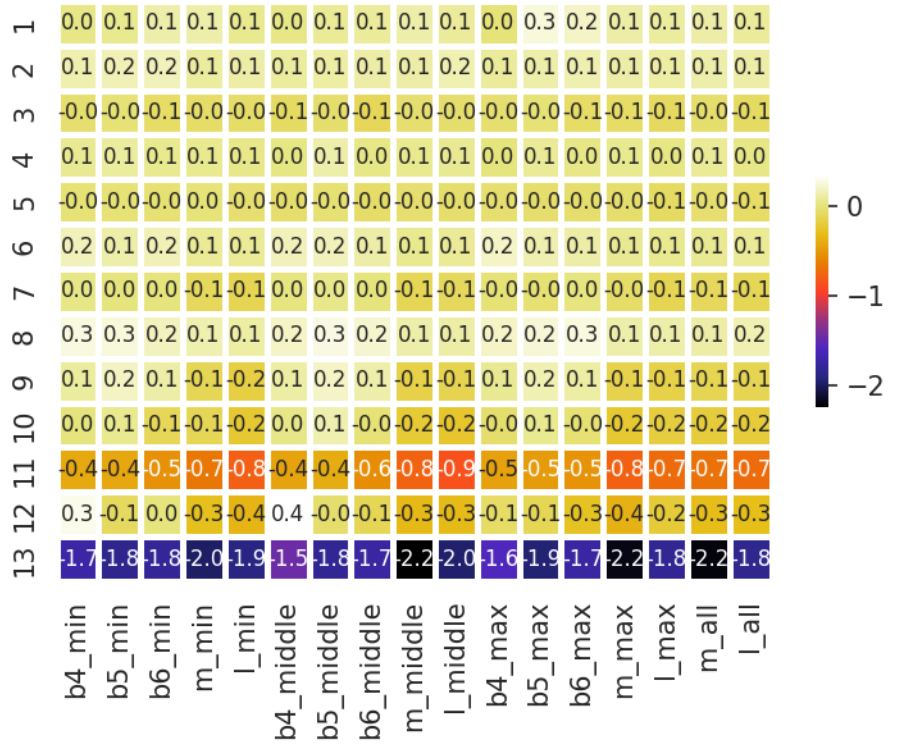


Figure 16. Standard deviation of residuals per age-group