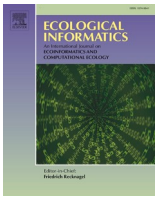
Ecological Informatics 78 (2023) 102325

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Age interpretation of cod otoliths using deep learning

Endre Moen [a,*], Rune Vabø [a], Szymon Smoliński [b], Côme Denechaud [a], Nils Olav Handegard [a], Ketil Malde [a,c]

[a] Institute of Marine Research, Bergen, Norway
[b] Department of Fisheries Resources, National Marine Fisheries Research Institute, Kołłątaja 1, 81-332 Gdynia, Poland
[c] Department of Informatics, University of Bergen, Norway

ARTICLE INFO

Keywords:
Otoliths
Aging
Machine learning
Deep learning
Images

ABSTRACT

Fish age estimation plays a crucial role in stock management and provides valuable information for biological studies. Fish age is typically estimated by experts that manually count annual increments in otoliths. This process is prone to age reader bias, which makes comparisons between readers and labs challenging, and requires considerable time and resources. In this study, we developed a machine learning framework for fish age prediction using 5150 images of otoliths from Barents Sea Atlantic cod (*Gadus morhua*) collected between 2012 and 2018. In contrast to previous studies that utilise models trained on otolith sections, we used images of broken otoliths that require no processing prior to imaging, and hence, could potentially facilitate at-sea age estimation. We trained convolutional neural networks (CNNs) based on two modern architectures (EfficientNetV1 and EfficientNetV2), which vary in model size (number of model parameters), and compared performance. Model average accuracy was 72.7% and mean-squared-error was 0.284 when compared with the human-read ages. The models' accuracy for one- and two-year-old individuals was over 90% and no systematic bias in the age predictions across age groups was detected. The best models were EfficientNet B4 and EfficientNet B6 using images taken with low exposure times. A maximum accuracy of 78.6% was achieved using an ensemble consisting of six models. Model predictions were also strongly correlated, limiting the utility of building large ensembles Model performance was compared to the results of an internal workshop where 100 independent images of broken otoliths were aged by a group of experts. Variations in percentage agreement between age classes showed a similar pattern (decreasing with age) in both CNN-based predictions and age estimates made by the expert group. While CNN-based percentage agreement was often lower than the expert estimates, it remained within or close to the range of percentage agreement observed across all readers. Our results demonstrate the potential of deep learning techniques for extracting age estimates from otolith images. When developing frameworks for age estimation using machine learning, we recommend EfficientNet B4 models are used as they are quicker to train than larger models and perform well. Ensemble approaches are also recommended if sufficient computational resources are available, as they can provide increased accuracy and lower variance of the predictions.

1. Introduction

Knowledge of fish age structure is central to fisheries science and modelling stock dynamics. It provides information on population growth and mortality and is one of the main criteria used for determining the health of exploited populations and monitoring the effects of selective fishing (Brunel and Piet, 2013; Hidalgo et al., 2011). Monitoring changes in the age distribution of a fish population can help to track substantial changes in population structure, such as the appearance of a particularly strong year-class (Reglero and Mosegaard, 2006), or the gradual truncation of older age classes as selective fishing mortality removes larger individuals (Siskey et al., 2016). Hard structures such as scales and otoliths, which form regular and temporally resolved growth increments, can be used to estimate fish age at an annual level (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,

* Corresponding author.
E-mail address: endre.moen@hi.no (E. Moen).

https://doi.org/10.1016/j.ecoinf.2023.102325
Received 12 May 2023; Received in revised form 2 October 2023; Accepted 2 October 2023
Available online 5 October 2023

requires specific expertise and is subject to uncertainties in both between-reader precision and "true" age accuracy (Francis and Campana, 2011). These biases can occur between readers both within and between otolith laboratories. Therefore, streamlining, scaling, and increasing the quality of age estimations can improve the reliability of evaluations of fish biology and consequently assessment of stock size and structure (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 and relatively untransparent otoliths that typically require time-consuming preparation (Denechaud et al., 2020; Smoliński et al., 2020). These routines vary between species and populations and institutes and range from a direct reading of broken otoliths under a magnifying glass, to embedding, thin sectioning, and finally imaging of the sections under a microscope. All otoliths read in Norway and Russia for the Northeast Arctic cod population are read on broken otoliths using a magnifying glass.

There has been a variety of methods proposed to automatically interpret otoliths from images, 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; Sigurdardóttir et al., 2023). One of the main advantages of automation is that the results are reproducible and consistent. Age predictions obtained from the automatic algorithm can be used e.g. for the quality control and identification of age reader biases within and between otolith laboratories (ICES, 2013).

1.1. Deep learning and image analysis

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

The main objective of this study was to develop a deep learning framework for automating the age estimation of Atlantic cod based on images of broken otoliths taken with constant illumination and three different exposures. We tested EfficientNetV1 and EfficientNetV2 architecture families using a range of model sizes from each, and we compared the performance of individual models and ensemble model runs. We also provide best practices and strategies for developing CNN frameworks for fish age predictions based on the images of otoliths. We further anticipate that this can serve as a baseline for the future development and operationalization of CNN models and the inclusion of ML-based otolith age interpretations during research expeditions and stock assessment surveys.

2. Method and materials

2.1. Data collection

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

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

2.2. Convolutional neural network architecture

Each CNN was trained using transfer learning by loading ImageNet weights. The training images were resized from 3744 × 5616 pixels to between 380 × 380 and 528 × 528 pixels depending on the architecture. The pixel values have a range between 0 and 255, which was normalized to between 0 and 1. Test set predictions were done on images resized to 380 × 380 and 384 × 384 pixels. To investigate the effect of exposure and orientation as presented in the image-taking protocol described in (Myers et al., 2019), we also trained on 9-channel images by stacking the three colour layers from each of the three images representing different lighting exposures. Using Timm (Wightman, 2019), the ImageNet weights were duplicated on the input layer to accommodate all 9 channels. The three images used were of dark, medium, and light exposure.

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

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

2.3. Implementation and training

EfficientNetV1 B4, B5, and B6 were imported and modified with TensorFlow (Abadi et al., 2016) and Keras (Chollet and others, 2018) software packages in Python. Computation was done using CUDA 11.1 and CuDNN with Nvidia (Nvidia Corp., Santa Clara, California) A6000

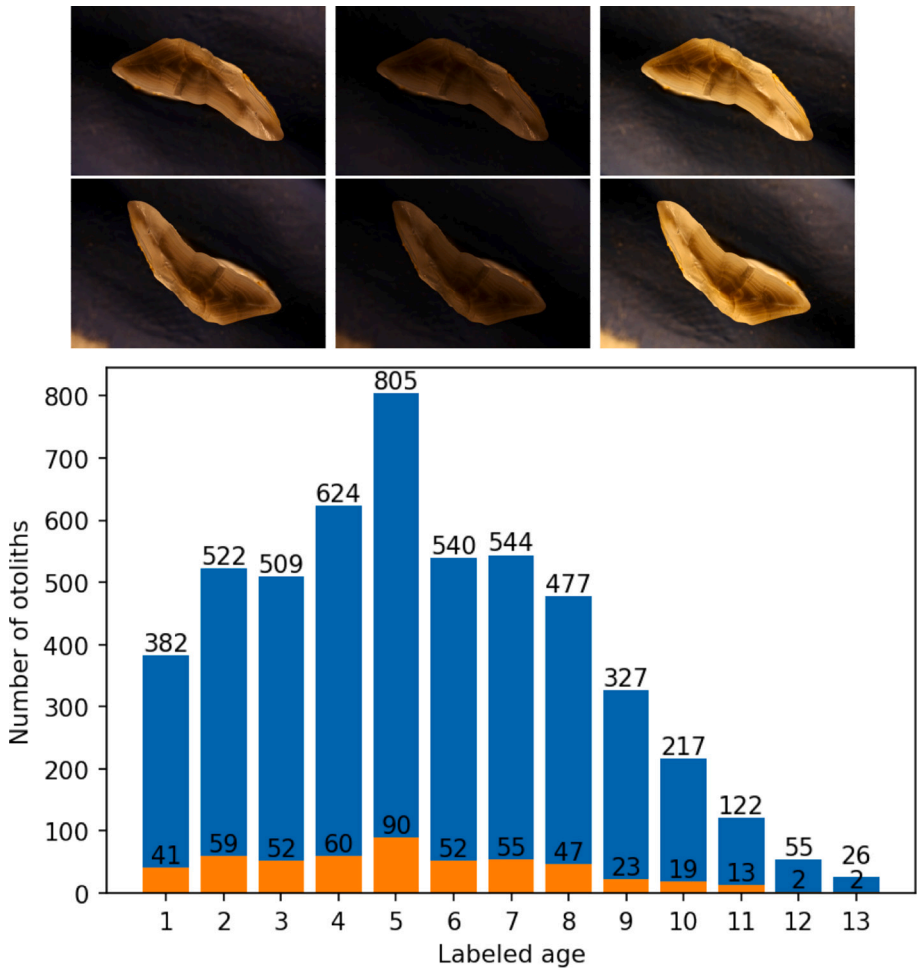


Fig. 1. Images of an otolith collected in 2016 from a 6 years old cod (top), taken with medium-, minimum- and max-exposures (upper row), then rotated 180° (lower row). The age distribution of the 5150 otoliths in the training (blue), and the 515 otoliths in the test (orange) set (bottom). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

accelerator card with 48 GB of GPU memory and P100 cards with 12 GB of GPU memory, EfficientNetV2 Medium, and Large were imported and modified with the PyTorch (Paszke et al., 2019) and Timm (Wightman, 2019) software packages. Computation was done on P100 and RTX 3090 with 24 GB of GPU memory. Pretrained weights for EfficientNet were available from Keras, and pre-trained weights for EfficientNetV2 were available from Timm. The models will be referred to as B4-Min, B4-Middle, B4-Max, B5-Min, Medium-min and so on by combining model name with image exposure.

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

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

The data set of 5150 otoliths was divided into a training set constituting 90% of the otolith images (4635 otoliths) and a test set of 10% (515 otoliths). To get the most out of a small data set we applied 10-fold cross-validation on the training set. The data set is divided into ten parts, and in each iteration (or"fold"), a different part is retained for validation, while the model is trained on the remaining nine parts. In other words, 10 different models were trained with a different set of 463 images used for validation in each fold, i.e. each data point participates in the validation set once and in the training set 9 times. Among the 10-fold models, the one with the best MSE was chosen. The best model parameters on the validation set were then used to predict the age on the test set, and the metric for accuracy and MSE were recorded. The test set is chosen at random, while the 10-fold split of the training set is chosen using a stratified k-fold 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.

2.4. Hyperparameters

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

EfficientNet used learning-rate with a weight decay scheduler, while

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

2.5. Ensemble learning with averaging

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

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

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

2.6. Correlations of predictions on the test set and clustering analysis

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

2.7. Comparison of CNN with human readers

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

3. Results

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

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

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

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

	EfficientNet V1			EfficientNet V2		
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**
MSE:light/CNN						
min	0.277	0.277	0.272	0.273	0.280	0.276
middle	0.285	0.273	**0.262**	0.278	0.275	0.275
max	0.291	0.359	0.305	0.289	0.286	0.306
9 channels	–	–	–	0.273	0.271	0.272
Mean	**0.284**	**0.303**	**0.280**	**0.278**	**0.278**	**0.284**
PA:light/CNN						
min	89.5	89.3	88.2	89.7	89.9	89.3
middle	88.2	89.5	90.9	91.1	87.8	89.5
max	87.6	90.5	88.0	89.5	90.3	89.2
9 channels	–	–	–	**91.3**	91.1	91.2
Mean	**88.1**	**89.8**	**89.0**	**90.4**	**89.8**	**89.6**

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

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

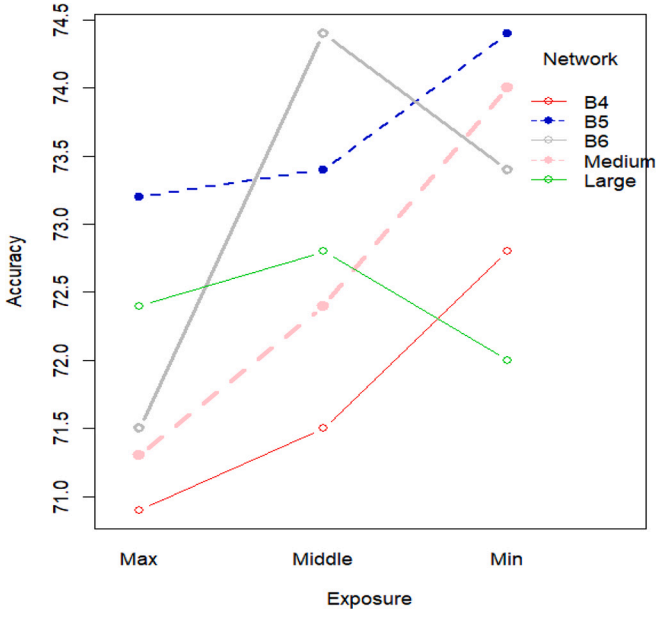


Fig. 2. Interaction plot of the 5 networks with image exposure on x-axis and ensemble accuracy on the y-axis. We see that under-exposed images perform better for all but the B6 and Large network.

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

3.1. Prediction by age class

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

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

3.2. Simple ensemble-average predictions

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

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

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

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

3.3. Outliers

Fig. 5 shows 4 images that were incorrectly classified with an error larger than 1 year after rounding. All the images with more than 1 year in prediction error are shown in supplementary information (Table 7), with comments by an expert on the most common mispredictions (Table 8). Large outliers occurred throughout all of the tested models and ensembles in small numbers. Most of those outliers were identified as visually challenging images with artifacts and/or low readability. For example, image 13 was overestimated in all B models, likely due to a clear zone in the inner core region that an expert reader would identify as a settlement false zone and ignore. Similarly, many outliers, such as images 270 and 369, showed multiple narrow false zones in the mid-section of the otolith that were likely to affect age determination. Alternatively, cases such as images 71 and 342 showed clear issues with age interpretation when the image deviates from the standard of the training set, such as when the exposure was changed drastically or when break lines interrupted the normal pattern of ring deposition. In one case (image 362), all models estimated the otolith to be 5 instead of 7 years old: upon visual investigation, the otolith was clearly 5 years old, and the initial age had likely been misread.

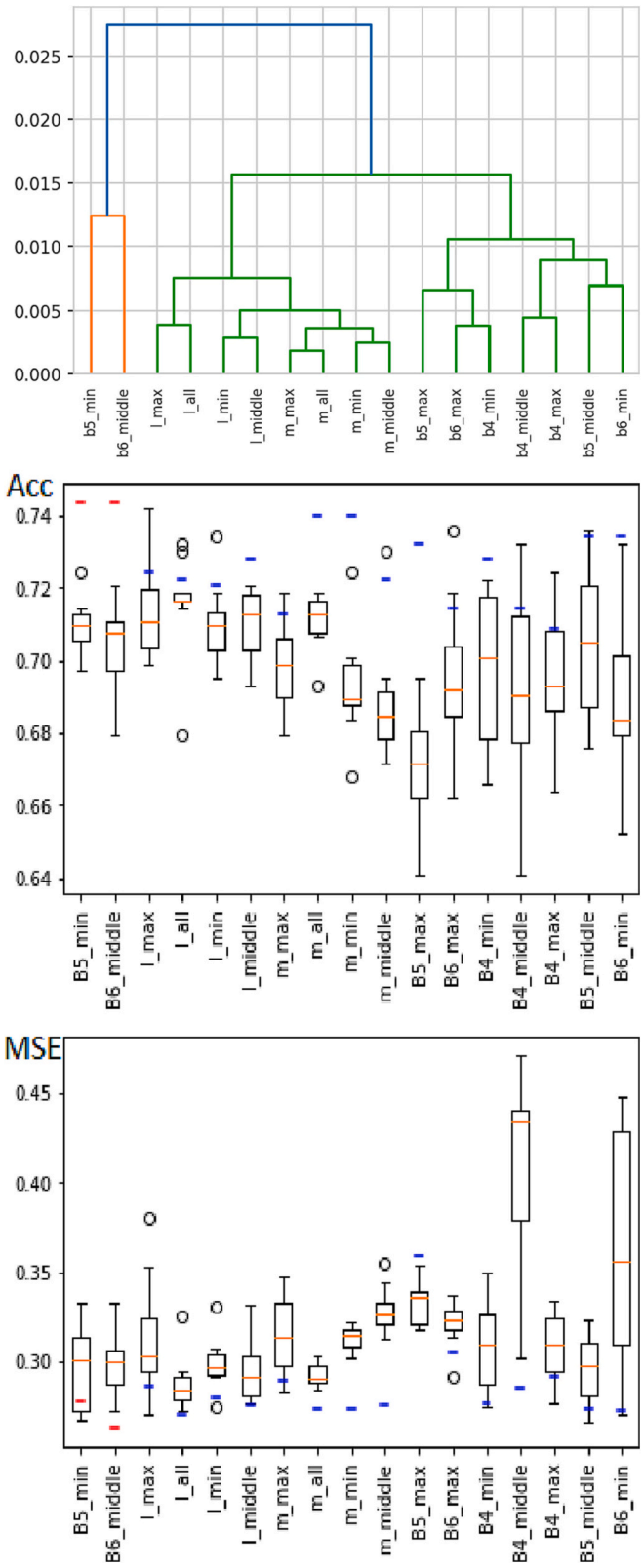


Fig. 3. Hierarchical clustering (HCA) on the correlation of predictions (top), a box-plot of accuracy score (middle), and MSE (bottom) of all the 17 models. In (middle) and (bottom), the blue line is ensemble-average prediction accuracy (or MSE) on the test set, the red lines are the two best ensemble-average predictions on the accuracy, and the orange lines are the mean of the 10-fold predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

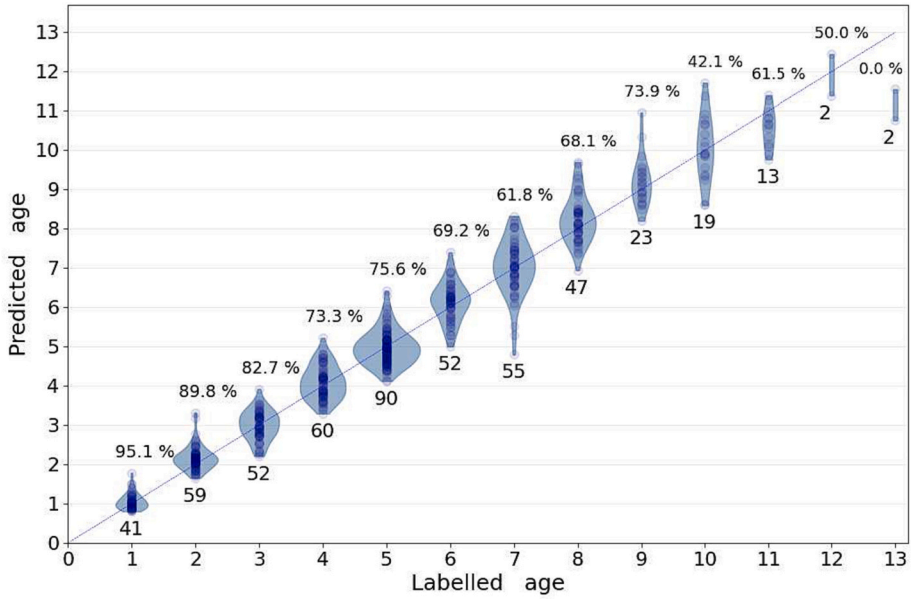


Fig. 4. Violin plot of predicted age from model B5-min with accuracy of 74.4%. Above each age is the accuracy, and below is the total number of images in the test set of that age class.

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

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

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

We also observed that the number of outliers did not correlate with model performance. E.g., 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 outliers with only 6 mispredictions.

3.4. Correlation of predictions and cluster analysis

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

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

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

3.5. Comparison of CNN with human readers

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

4. Discussion

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

4.1. Accuracy across different age groups

The age of the younger individuals was predicted with greater accuracy using the CNN models than those of older individuals. Thus, the CNN appears to be particularly competent at aging cod otoliths of younger age classes. This is also typical for expert readers who generally show the greatest accuracy for the youngest age classes which have fewer and clearer rings (Campana, 2001). However, the reasons both humans and CNNs find the age of younger individuals easier to predict may not be the same. Human expert readers use various visual cues, prior knowledge, and background information to determine fish age, such as comparing ring counts on multiple axes and having intrinsic knowledge of the periodicity of opaque and translucent zones for a given species. In younger fish, the increments are usually wider and more

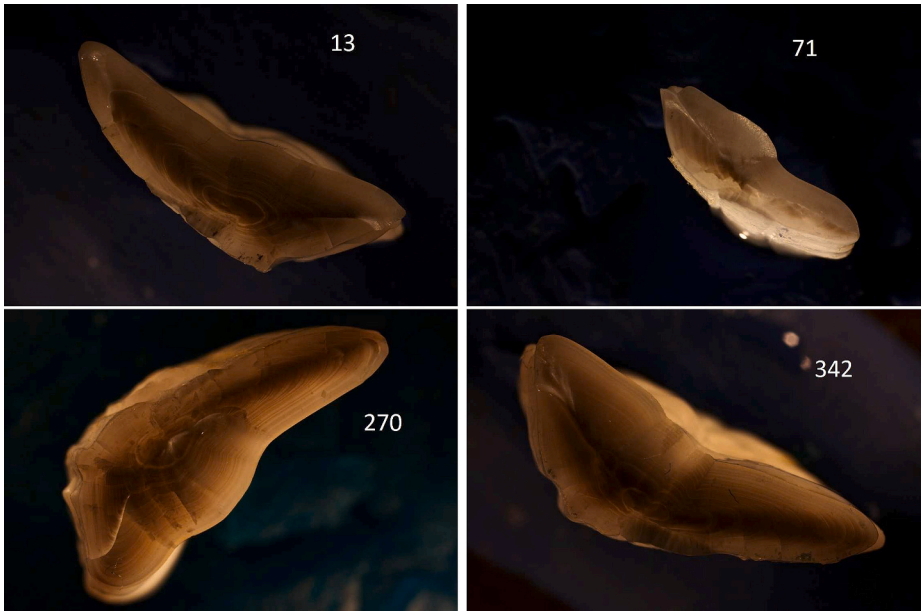


Fig. 5. Example outlier images with index 13, 71, 279, 342 from the test-set were mispredicted by between 25% and 100% of the models.

clearly separated as fish -and consequently otolith- growth rates are maximal prior to maturity. Fish of age 1 are small and have comparatively small otoliths with a straightforward ring pattern made of one single finished opaque and translucent zone, and expert readers are unlikely to disagree on its interpretation. On the other hand, a CNN architecture as used here identifies hierarchical patterns on different scales of the image from which it derives a value in the range of those provided in the training set. This means that unless specifically forced to do so, the algorithm may seek and interpret visual clues other than the rings human readers are trained to use. A possible explanation for the higher prediction accuracy of younger fish is that age is related to the area the otolith covers relative to the total image size. Because the same camera settings were used, all images had the same dimension and calibration. For a species with moderately large adults such as Atlantic cod (Froese and Pauly, 2022), the otoliths will grow in size significantly faster during the first years and then slow down with approaching sexual maturity. As fish get older different growth trajectories will then lead to greater overlap in otolith sizes across different ages. It is therefore possible that the CNNs are not counting growth zones as human expert readers would, but rather that they synthesize all available patterns in the image to find recurring age-related characteristics evident in the training set. The size of the area that the otoliths cover against the more uniform black background might for example be a very simple feature picked up by the CNNs with high predictive power for the youngest fish., while the higher inter-individual variability and greater size overlap at older ages would affect the predictive accuracy of CNNs.

The hypothesis that CNNs exploit other information than the growth zones is consistent with the findings of an earlier study where network activations inside a CNN were explored for images of Greenland halibut otoliths (Ordonez et al., 2020). Visualisation techniques were used to reveal the relative importance of attributes such as shape, inner structure, and size of the otoliths using activation maps. Importantly, the authors found that the CNN utilized information in pixels corresponding to annual increments to only a small extent. To explore this possibility, we attempted to train a network using otolith silhouettes only, e.g. images where all internal structures where erased. We could not achieve acceptable performance of the CNN models during these initial tests.

4.2. Sub-year agreement between models

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

4.3. Importance of training set size relative to model performance

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

4.4. The effect of image size

The high-resolution 3744×5616 cod otolith images were scaled down to between 380×380 and 528×528 pixels to match the requirements of the different EfficientNet architectures. This reduction in resolution may have affected the readability of finer-scale visual features such as growth rings. The fixed camera setup resulted in the background constituting a large proportion of the images, especially for smaller otoliths from the youngest individuals. This is especially true due to the curved or oval shape of the otolith, as a compressed image will not only have less pixels to work with but will also have a comparatively more important fraction of black background, which is effectively useless for age interpretation. Otoliths of other fish species like red mullet which have a more circular shape may be less sensitive to this problem

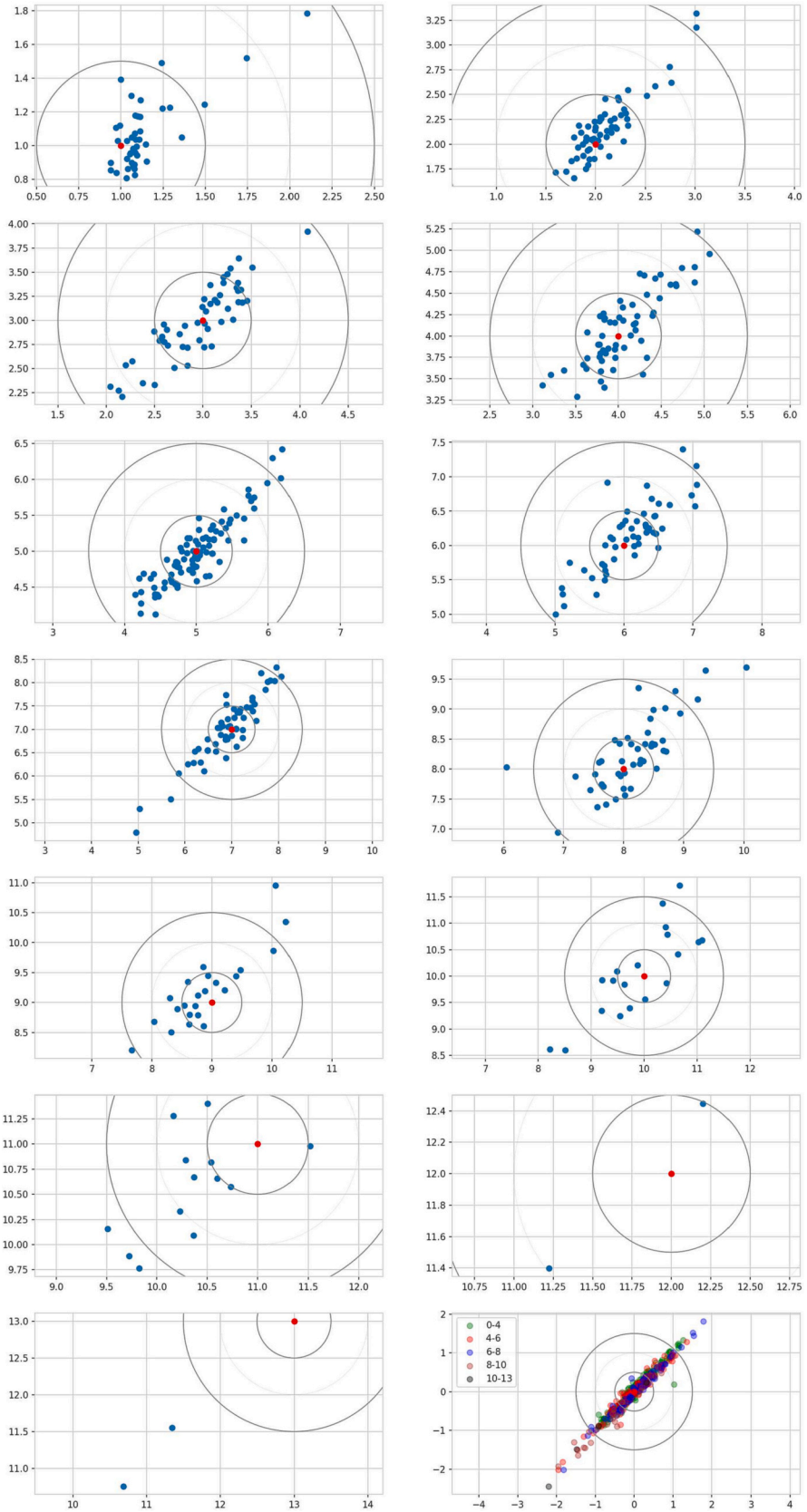


Fig. 6. Comparison of age estimates predicted by Medium-min (x axis, years) and B5-min (y axis, years) as age-specific scatter plots, and in aggregate for all age groups in the bottom right panel. The circles show age differences of 0.5 and 1.5 years.

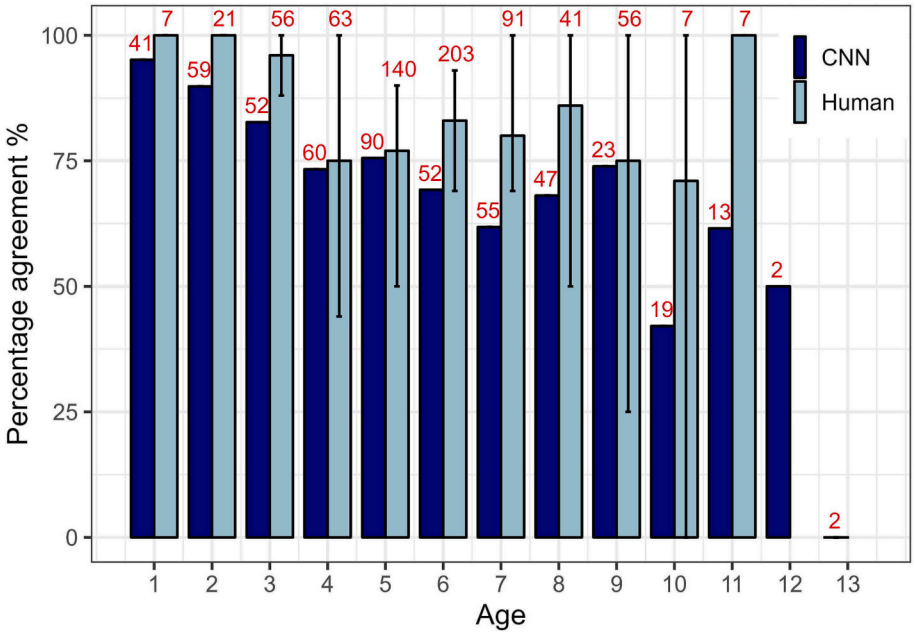


Fig. 7. Comparison of mean percentage agreement within each age class for two sets of otoliths: the CNN-predictions on the test set (black); an internal age reading of 100 cod otoliths involving 7 readers (gray). Numbers indicate the total number of readings for each age class (with 1 reading per otolith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

(Politikos et al., 2021). Improvements might therefore be made by first isolating the otoliths from their background and cropping the image accordingly, before training the network on the information contained exclusively within the area of interest. This would also limit information loss from image compression.

4.5. Outliers and transparency

The CNNs trained in this study produced predictions of fish age that were generally within less than a year off the labeled values. It is noteworthy that predicted ages were similar across different models and errors of more than a year were only seen in 2% of the predictions made using the test set. Closer inspection revealed that such errors were often caused by otolith images with poor readability, in particular drastic changes in exposure or visual damages and interruptions on the reading axis.

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

4.6. Effects of image exposure on predictive power

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

4.7. Effects of 9-channel composite images and architecture size

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

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

4.8. Utilizing model ensembles

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

4.9. Comparison of CNN with human readers

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

4.10. Resource efficiency

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

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

CNNs can also be applied without high additional cost or even be incorporated into routine protocols and provide additional value e.g., reading consistency check, time-drift evaluations, inter-reader comparisons (how much each reader is 'off' when compared to CNN predictions, even if not compared with the same otolith samples), etc. Also, networks are comparatively easier to scale up than the number of human expert readers, especially when analysing huge datasets, as any increase in ensemble size (number of networks) only requires an increase in computer resources.

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

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

4.11. Conclusion

Our results demonstrate that deep learning techniques have huge potential in extracting age information from otolith images. Standard model architectures trained on sufficient training data specific to the use case can accurately predict age from images of broken otoliths. We believe that carefully trained CNNs could become a major component in procedures that require minimal processing and could be able to produce near at sea age estimates. In addition, algorithmic age estimates could serve as a useful reference for evaluating age reader biases within and between otolith laboratories.

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

Data availability

Data will be made available on request.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2023.102325.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al., 2016. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems arXiv preprint arXiv:1603.04467.

Albuquerque, C.Q., Lopes, L.C.S., Jaureguizar, A.J., Condini, M.V., 2019. The visual quality of annual growth increments in fish otoliths increases with latitude. Fish. Res. 220, 105351.
Beamish, R.J., McFarlane, G.A., 1995. A discussion of the importance of aging errors, and an application to walleye Pollock: the world's largest fishery. In: Recent Developments in Fish Otolith Research. University of South Carolina Press, Columbia, S.C., pp. 545–565
Brunel, T., Piet, G.J., 2013. Is age structure a relevant criterion for the health of fish stocks? ICES J. Mar. Sci. 70, 270–283.
Campana, S., 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. J. Fish Biol. 59 (2), 197–242.
Carbonara, P., Follesa, M.C., 2019. Handbook on fish age determination: a Mediterranean experience. In: General Fisheries Commission for the Mediterranean. Studies and Reviews, 98, pp. 1–179.
Chollet, F., others, 2018. Keras 2.1.3. https://github.com/fchollet/keras.
Denechaud, C., Smoliński, S., Geffen, A.J., Godiksen, J.A., Campana, S.E., 2020. A century of fish growth in relation to climate change, population dynamics and exploitation. Glob. Chang. Biol. 26 (10), 5661–5678.
Fisher, M., Hunter, E., 2018. Digital imaging techniques in otolith data capture, analysis and interpretation. Mar. Ecol. Prog. Ser. 598, 213–231.
Francis, R.C., Campana, S.E., 2011. Inferring age from otolith measurements: a review and a new approach. In: Canadian Journal of Fisheries and Aquatic Sciences. NRC Research Press Ottawa, Canada. https://doi.org/10.1139/f04-063 (Accessed 3 February 2022).
Froese, R., Pauly, D., 2022. Fishbase.
Hidalgo, M., Rouyer, T., Molinero, J.C., Massutí, E., Moranta, J., Guijarro, B., Stenseth, N.C., 2011. Synergistic effects of fishing-induced demographic changes and climate variation on fish population dynamics. Mar. Ecol. Prog. Ser. 426, 1–12.
Høie, H., Millner, R.S., McCully, S., Nedreaas, K.H., Pilling, G.M., Skadal, J., 2009. Latitudinal differences in the timing of otolith growth: a comparison between the barents sea and southern north sea. Fish. Res. 96, 319–322.
ICES, 2013. Report of the second workshop of National Age Readings Coordinators (WKNARC2). In: Report of the Second Workshop of National Age Readings Coordinators (WKNARC2). ICES, Copenhagen.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, 25. Curran Associates, Inc, pp. 1097–1105.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Loshchilov, I., Hutter, F., 2016. Sgdr: stochastic gradient descent with warm restarts. Neurips.
Mahé, K., 2009. Project no. 044132. Automated FISh Ageing (AFISA): Final Activity Report.
Moen, E., Handegard, N.O., Allken, V., Albert, O.T., Harbitz, A., Malde, K., 2018. Automatic interpretation of otoliths using deep learning. PLoS One.
Myers, S., Thorsen, A., Godiksen, J., Malde, K., Handegard, N., 2019. An efficient protocol and data set for automated otolith image analysis. GeoSci. Data J.
Ordonez, A., Eikvil, L., Salberg, A.-B., Harbitz, A., Murray, S.M., Kampffmeyer, M.C., 2020. Explaining decisions of deep neural networks used for fish age prediction. PLoS One 15 (6), e0235013.
Panfili, J., de Pontual, H., Troadec, H., Wrigh, P.J., 2002. Manual of Fish Sclerochronology. *https://archimer.ifremer.fr/doc/00017/12801/* (Accessed 3 February 2022).
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., 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., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc, pp. 8024–8035.
Politikos, D.V., Petasis, G., Chatzispyrou, A., Mytilineou, C., Anastasopoulou, A., 2021. Automating fish age estimation combining otolith images and deep learning: the role of multitask learning. Fish. Res. 242, 106033.
Ragonese, S., 2018. Methuselah or butterfly? When fish age estimates and validations tell different stories. In: The Case of the European Hake (Merluccius Merluccius l. 1758) in the Mediterranean Sea. *https://www.researchgate.net/profile/Sergio-Ragonese/publication/328191704_Methuselah_or_butterfly/links/5bbdca0445851572315bddec/Methuselah-or-butterfly.pdf*".
Reglero, P., Mosegaard, H., 2006. Onset of maturity and cohort composition at spawning of Baltic sprat sprattus sprattus on the basis of otolith macrostructure analysis. J. Fish Biol. 68, 1091–1106.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2014. Imagenet Large Scale Visual Recognition Challenge.
Sigurdardóttir, A.R., Sverrisson, D., Jónsdóttir, A., Gudjónsdóttir, M., Elvarsson, B.D., Einarsson, H., 2023. Otolith age determination with a simple computer vision based few-shot learning method. Eco. Inform. 76, 102046.
Siskey, M.R., Wilberg, M.J., Allman, R.J., Barnett, B.K., 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 J. Mar. Sci. 73, 2518–2528.
Smoliński, S., Deplanque-Lasserre, J., Hjörleifsson, E., Geffen, A.J., Godiksen, J.A., Campana, S.E., 2020. Century-long cod otolith biochronology reveals individual growth plasticity in response to temperature. Sci. Rep. 10 (1), 1–13.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the inception architecture for computer vision. CoRR abs/1512.00567.
Tan, M., Le, Q.V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. CoRR abs/1905.11946.
Tan, Mingxing, Q. V. L, 2021. Efficientnetv2: Smaller models and faster training. CoRR abs/2104.00298.
Tyler, A.V., Beamish, R.J., McFarlane, G.A., 1989. Implications of age determination errors to yield estimates. ICES J. Mar. Sci. 108, 27–35.
Vabø, R., Moen, E., Smoliński, S., Husebø, Åse, Handegard, N.O., Malde, K., 2021. Automatic interpretation of salmon scales using deep learning. Ecol. Inform. 63, 101322 (2021).
Wightman, R., 2019. Pytorch Image Models. https://github.com/rwightman/pytorch-image-models.
Xu, B., Wang, N., Chen, T., Li, M., 2015. Empirical evaluation of rectified activations in convolutional network. CoRR abs/1505.00853.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, 27. Curran Associates, Inc, pp. 3320–3328.