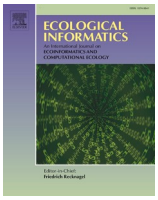
Ecological Informatics 76 (2023) 102046

Contents lists available at ScienceDirect

Ecological Informatics

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



Otolith age determination with a simple computer vision based few-shot learning method

Andrea Rakel Sigurðardóttir [a,c,*], Þór Sverrisson [a], Aðalbjörg Jónsdóttir [b], María Gudjónsdóttir [c,d], Bjarki Þór Elvarsson [b], Hafsteinn Einarsson [a]

[a] University of Iceland, Faculty of Computer Science, Iceland
[b] Marine and Freshwater Research Institute, Iceland
[c] University of Iceland, Faculty of Food Science and Nutrition, Iceland
[d] Matis, Food and Biotech R&D, Iceland

ARTICLE INFO

Keywords:
Otoliths
Fish age estimation
Few-shot learning
Deep-learning
Image analysis

ABSTRACT

In this study, we propose a computer vision-based few-shot learning method for otolith age determination in European plaice, Atlantic cod, Greenland halibut, and haddock. Our method outperforms prior state-of-the-art approaches, and is based on a vision encoder from CLIP as a feature extractor, which is used to train shallow models. The method is computationally efficient, as it does not require fine-tuning of deep networks, and is also data efficient, as it performs better than fine-tuning on the same data. Our results suggest that in some cases, our method can achieve the same performance as state-of-the-art finetuning approaches with up to three times less training data.

1. Introduction

To support sustainable fishing practices, it is important to evaluate the status of fish stocks and their productivity to gain a deeper understanding of population dynamics. This evaluation involves estimating the age distribution of the fish population (Carbonara and Follesa, 2019). That information is used to predict the size of future generations and determine fishing quotas. An established method used today for age determination of fish is based on sclerochronology. In sclerochronology, the aim is to reconstruct the past history of a living organism through the study of calcified structures (Panfili et al., 2002). In our setting, this involves the manual reading of annual rings present in fish otoliths since they are incremental structures related to fish growth (Wang et al., 2019). Manual age determination is time-consuming and costly, as it is labour intensive and requires otolith reading expertise. Automating the task of otolith age determination is therefore of interest to better support sustainable fishing practices. Ideally, an efficient, precise and cost-effective method could aid experts in otolith readings.

Computer vision (CV) approaches, such as deep learning, have been used for otolith age determination in many species. For example, Moen et al. (2018) fine-tuned an Inception-V3 model to automatically read otolith images from Greenland halibut (*Reinhardtius hippoglos-soides*), achieving 29% accuracy. The accuracy improved by additional 38% when an error of 1 year was allowed in the comparison to the annotated reader age. Moore et al. (2019) used a similar approach to fine-tune an Inception-V3 model for Snapper (*Chrysophrys auratus*) and Hoki (*Macruronus novaezelandiae*) otoliths, achieving 47% and 41% accuracy, respectively. Politikos et al. (2021) used a fine-tuned Inception-V3 model with multitask learning to predict both the age and length of red mullet (*Mullus barbatus*) otoliths, achieving 64% accuracy when fine-tuned only for age prediction, and 69% accuracy when fine-tuned for both tasks. Martinsen et al. (2022) used a machine learning model called Xception to determine the age of Greenland halibut based on otolith images, achieving 24.2% accuracy when trained on male and female fish, and 64.4% accuracy when a margin of error of ± 1 was allowed. The study also included a linear regression model based on fish length and sex, which had 19.8% accuracy without a margin of error, and 55.7% accuracy with a ± 1 margin of error.

Previous work has shown that deep-learning-based approaches are effective for automated otolith age determination. Many of these approaches involve fine-tuning the Inception-V3 model, a convolutional neural network developed by Google (Szegedy et al., 2015a). Finetuning such a model requires a large dataset of annotated examples, which must be labeled manually (for a background text on deep-learning, see

* Corresponding author at: University of Iceland, Faculty of Computer Science, Iceland.
E-mail address: ars59@hi.is (A.R. Sigurðardóttir).

https://doi.org/10.1016/j.ecoinf.2023.102046
Received 13 December 2022; Received in revised form 28 February 2023; Accepted 1 March 2023
Available online 5 March 2023
1574-9541/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Goodfellow et al., 2016). The performance of the model is determined by the number and quality of these examples. Typically, the more examples used, the better the model performs. However, since these models can require thousands of examples to reach their full potential, it is important to study methods that require fewer annotated examples to achieve the same level of performance.

Manual labeling of otoliths for age determination can be expensive, which can limit the amount of labeled data available for machine learning projects in this field. Few-shot learning is a method that aims to overcome this challenge by training machine learning models on a small number of labeled sample images, with the goal of achieving good performance on age prediction tasks. This approach can help to make up for the lack of labeled data available for training. Few-shot learning has been used in CV projects such as character recognition (Shaffi and Hajamohideen, 2021), image classification (Radford et al., 2021), and object detection (Xu et al., 2016), but has not yet been applied to otolith age prediction.

Recent progress in few-shot learning has been based on transformers. These models were first applied successfully in natural language processing (NLP), and are used in popular language models such as Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019). They have also been implemented in vision models, such as the Vision Transformer (ViT) (Dosovitskiy et al., 2023). The transformer architecture, introduced by Vaswani et al., 2017, uses an attention mechanism such that the model can learn what parts of the input it should attend to. This allows it to outperform state-of-the-art models in NLP tasks without using recurrence in the model architecture. Dosovitskiy et al., 2023 showed that convolutions are not necessary for state-of-the-art performance in image classification tasks, and that transformers can perform well on these tasks when applied to image patches.

For few-shot learning, CLIP (Contrastive Language Image Pre-training) is a combination of two transformer models that was recently introduced. It consists of a text encoder and an image encoder, and is trained using contrastive learning to allow the image encoder to learn visual concepts through natural language supervision (Radford et al., 2021). The image encoder maps an image to a vector that can be considered a robust semantic feature representation of the image, which can transfer to diverse tasks and often perform competitively with fully supervised baselines in a zero-shot fashion, without the need for fine-tuning. CLIP was trained on 400 million images from the web along with their captions. While these models have shown impressive performance, they are not perfect and have, for example, shown limited zero-shot performance in detecting tumors on x-ray images. We aim to investigate whether CLIP models are feasible for otolith age determination, and whether they can achieve higher performance than previous approaches with less labeled training data.

In this study, we propose a few-shot learning approach for otolith age determination in European plaice, Atlantic cod, Greenland halibut, and haddock. We use the vision encoder from CLIP as a feature extractor, and train linear and multiclass models using these features. Our approach is computationally efficient, as it does not require fine-tuning of deep networks, and is also data efficient, as it achieves better performance than fine-tuning on the same data.

2. Methods

2.1. Data acquisition

For this work we use otoliths from European plaice (*Pleuronectes platessa*), Atlantic cod (*Gadus morhua*), Greenland halibut (*Reinhardtius hippoglossoides*), and haddock (*Melanogrammus aeglefinus*) (see Table 1). The otoliths were collected and annotated by the marine and freshwater research institute (MFRI) in Iceland. The plaice otoliths were also provided with two variables relevant for age determination. A length measurement that can reflect a fish's age and the quarter in which the fish was collected, since that variable affects the age reading (Otolith readers assume that fishes are born at the beginning of the year). Greenland halibut otoliths were also provided with two relevant variables, sex and length measurements. Fig. 1 shows example otolith images of the species studied.

Table 1
Overview of the data used in this study.

Species	# Samples	# Annotators	Other features
European plaice (*Pleuronectes platessa*)	1000	3	Length, quarter
Greenland halibut (*Reinhardtius hippoglossoides*)	4821	1	Length, sex
Haddock (*Melanogrammus aeglefinus*)	3687	1	–
Atlantic cod (*Gadus morhua*)	1170	1	–

The Atlantic cod and haddock were collected during the MFRI's groundfish survey that is performed in the spring each year (Sólmundsson et al., 2010). The Greenland halibut otoliths were collected in groundfish surveys from the fall of 2014 to 2020. Finally, the plaice otoliths were collected as random samples from catches around Iceland.

The samples of the Greenland halibut were imaged using a Leica DFC295 camera with a resolution of 3.1 mp (2048 × 1536 px). The Atlantic cod was imaged using a Leica IC80 HD camera (2048 × 1536 px). The haddock was imaged with a Leica IC90 E camera (3648 × 2736 px). Unfortunately, the type of camera used for the European plaice otoliths was not documented in the collection process, but the images were 1280 × 960 px.

Age determination is performed by counting the ring patterns around the center of the otolith (Panfili et al., 2002). The distance between rings decreases with distance from the center as it reflects a lower growth rate. This reduced distance can make it challenging to determine the age of older fish. The mechanisms underlying the annual periodicities in otoliths are currently not understood but the biological and structural differences have been described (Katayama, 2018). Mounting evidence indicates that opacity changes in otoliths can be attributed to factors that influence the metabolic rate of fish such as ambient temperature changes and variations in food availability (Grønkjær, 2016). We show examples of plaice otoliths for different ages in Fig. 2. We note that the age is not only determined by the number of rings but also by the time at which the specimen was collected. The otolith readers assume that the fish are born at the beginning of the year and therefore the last ring is not counted towards the age if the specimen was collected in the fourth quarter.

2.2. Age determination

2.2.1. Few-shot learning approach

We explore several models for the age determination task using a transfer learning approach. In particular, we compare state-of-the-art multi-modal models as feature extractors with finetuning pre-trained models. An overview of the approach is shown in Fig. 3.

We use a pre-trained CLIP (Radford et al., 2021) image encoder for feature extraction along with a simple ridge regression or a multiclass support vector classifier. For the sake of comparison, we fine-tune a pre-trained visual transformer, as well as ResNet-50 and Inception-V3 which are deep convolutional neural networks.

For a multiclass classification approach, we use a Support Vector Classifier (SVC) using a one-vs-one scheme[1] with a linear kernel. For each binary classification problem, SVC finds a weight vector $\boldsymbol{w}$ and a bias term b by solving:

[1] The problem is set up as a binary classification problem for each pair of classes.

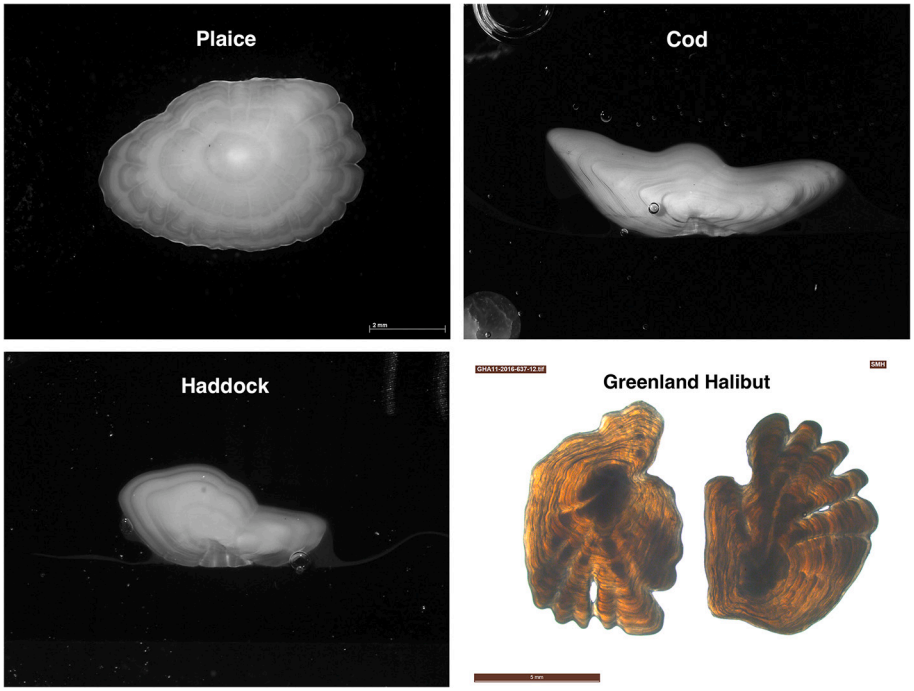


Fig. 1. Example images of different species from the dataset. The examples show otoliths of age 4 for European plaice, Atlantic cod, haddock and Greenland halibut.

$$\min_{w,b,\zeta} \frac{1}{2} \| w\|^2 + C\sum_{i=1}^{n} \zeta_i$$

subject to:

$$y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) \geq 1 - \zeta_i, \quad \zeta_i \geq 0$$

where $\boldsymbol{x}_i$ are the input vectors and $\boldsymbol{y} \in \{-1,1\}^n$ is the label vector. The penalty term C works as an inverse regularization parameter where ζ_i represents the distance of sample i from its correct margin boundary.

For linear regression, we use ridge regression where regularization is given by the L2-norm. We solve:

$$\min_{w} \| Xw - y\|_2^2 + \alpha \| w\|_2^2, \alpha \geq 0$$

where X is the input matrix and y the corresponding label vector. The regularization parameter α controls the strength of the L2 term.

For SVC and ridge regression, we use the implementation from Scikit-learn (Pedregosa et al., 2011).

2.2.2. Fine-tuning models

We compare fine-tuning of three models.

Vision Transformer (ViT) is a deep learning model for vision processing tasks that extends the original transformer architecture (Vaswani et al., 2017) designed for natural language processing applications (Dosovitskiy et al., 2023). For ViT, an image is split into patches which are fed into the transformer encoder, using attention to learn what parts of the image input are relevant.

ResNet-50 is a convolutional neural network model that uses skip connections, which made it possible to get good performance with deeper models than was previously possible (He et al., 2015). Training deep neural networks can result in exploding gradients, and skip connections were introduced to ameliorate that problem. ResNet-50 was chosen for comparison as it is a proven CV classification model. It is commonly used and provides a well-performing baseline.

Inception-V3 is a convolutional neural network used for image classification, that has an auxiliary classifier that acts as a regularizer (Szegedy et al., 2015b). The Inception-V3 architecture is built on previous Inception models, with the aim of making the V3 computationally more efficient than previous models. Inception-V3 was chosen for comparison with the other models in this study, as it has been the most popular CNN to use in previous publications on deep-learning-based automatic otolith age determination (Moen et al., 2018; Moore et al., 2019; Politikos et al., 2021).

2.2.3. Training

For the experiment, the plaice otoliths images were split 10 times into a train (65%), a validation (15%) and a test (20%) set. All splits were performed using the fish age for stratification. The models were trained on each split and the results reported are the average over all ten experiments. The validation set was only used for selecting the best model during fine-tuning. Using the ViT-L/14@336px image encoder, every otolith image was transformed into a one dimensional vector. It may be noted that downsampling and center cropping is built into the CLIP model. The length and the quarter parameters were appended to the end of the image vector. The quarters were one-hot encoded as a four-dimensional vector and the length was scaled by a factor of 0.01.

We used the Scikit-learn (Pedregosa et al., 2011) implementations of SVC and ridge regression. Regularization parameters for both methods were chosen by conducting a grid search using cross-validation on the training set. For SVC, we explored a C value in the interval [0.001, 1] (see Fig. C.2 in the Appendix C. Based on that result, the hyperparameter value chosen was $C = 0.1$. For ridge regression, we explored α values in the interval [0.1, 19.6] and concluded that the performance of the model is robust to the choice of the hyperparameter when it ranges from 5 to 15 (see Fig. C.1 in the Appendix C). Based on that result we chose to use $\alpha = 6.0$ for our experiments.

For fine-tuning, we used the HuggingFace implementations of ViT ("google/vit-hugepatch14–224-in21k") and ResNet ("microsoft/resnet-50") but the Keras implementation of Inception-V3. All fine-tuning models had been pre-trained on the ImageNet (Deng et al., 2009) dataset.

We replaced the classifier head on all models so that it received the output of the models along with the additional parameters, quarters and length. We also tried adding additional dropout and dense layers at the end of the Inception model as for the DeepOtolith model Politikos et al. (2021). To differentiate we use *Deep Otolith Inception-V3* for the model

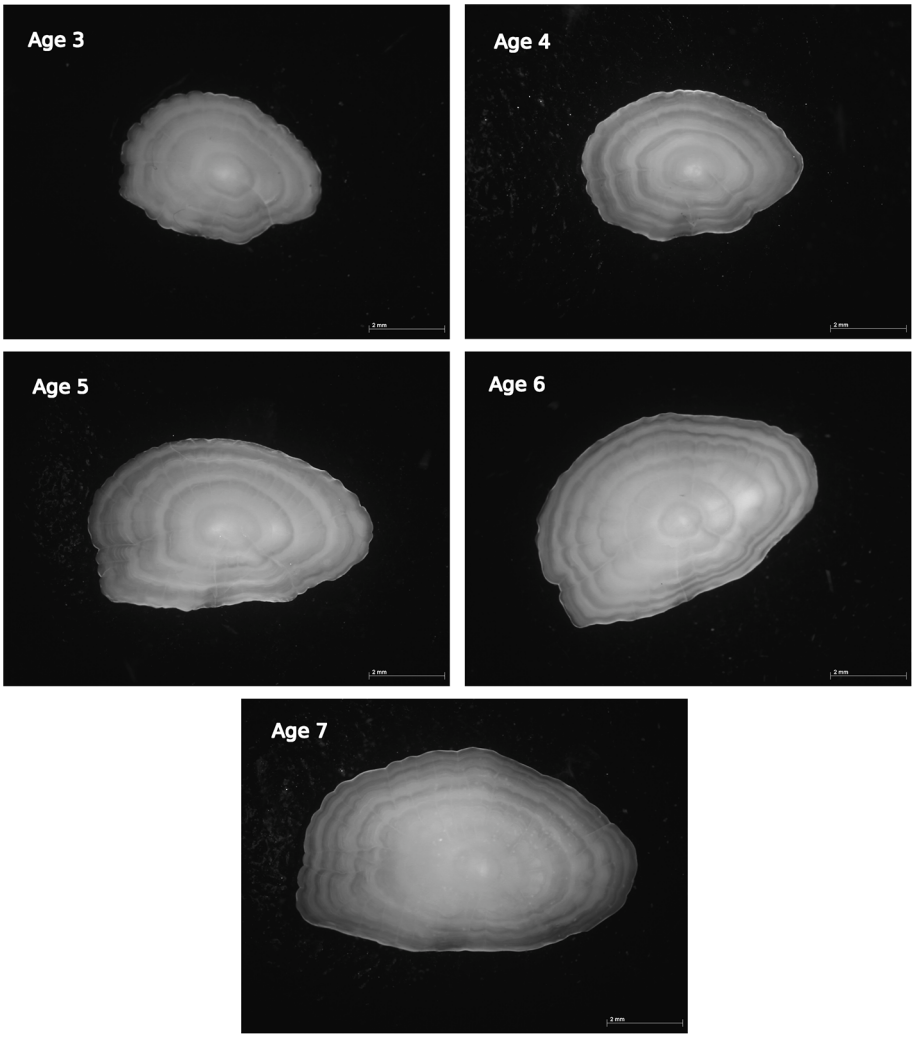


Fig. 2. Example images of different ages from the plaice otolith dataset. All annotators agreed on the age of these otoliths.

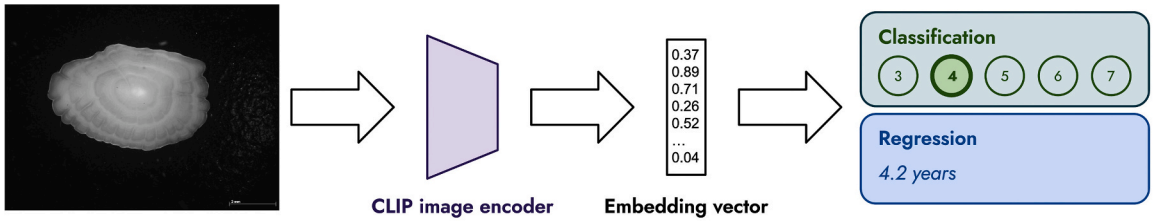


Fig. 3. Overview of the method. Images are passed through a CLIP image encoder and the resulting embedding vectors are used as input to classification and regression models to predict the otolith age.

with the added layers and *Vanilla Inception-V3* for the one without any additional layers.

The Inception-V3 model accepts images resized to (299, 299) pixels but ViT and ResNet resize to (224, 224) pixels. Dataset imbalance was dealt with by including class weights in the cross-entropy loss function. For the training we used AdamW optimizer for Hugging Face models, which is the default optimizer in the transformers library by HuggingFace. For Inception-V3, we used Adam since that was the optimizer used in previous work (and AdamW is not available in Keras). We used a batch size of 16 examples and the learning rate was set to 10^{-4} for ResNet but 10^{-5} for Vanilla Inception-V3 and ViT. For Deep Otolith Inception we used a learning rate of 4×10^{-4} as suggested by Politikos et al. (2021). For every fine-tuning experiment, the model with the smallest validation loss was used for evaluation on the test set.

2.3. Performance

For the evaluation of the performance of the models, we reported accuracy, precision, recall, and F1-score. The performance metrics are reported separately for each defined class. Averages over all classes are reported as well. The Root Mean Squared Error (RMSE) was used as an alternative performance metric.

By counting the occurrences of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) we can define the accuracy, precision, recall and F1-score of the models as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 - score = 2\,\frac{Precision \times Recall}{Precision + Recall},$$

and the RMSE is defined as follows:

$$\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}},$$

where $\hat{y}_i$ represents the predicted age of the i-th fish and y_i represents its annotated age. The predictions of the CLIP regression model were rounded to the nearest integer for a fair evaluation and comparison to the other models. However, when RMSE was computed, the output from the CLIP-regression model was not rounded. Performance metrics are and reported using Scikit-learn (Pedregosa et al., 2011) implementations.

3. Results

3.1. Otolith age distribution and characteristics

The age of the plaice otoliths was recorded independently by three readers as a discrete value with numbers ranging from 3 to 11. Most of the measurements ranged from 3 to 7 as shown in Fig. 4, so for the study we merged >7 age groups with 7. Reader agreement was high with 82.0% of cases having a perfect agreement, 17.2% with one disagreement, and 0.8% where all annotators disagreed. Fig. 5 shows examples where readers disagreed. Due to the high agreement, we set the median of the three age annotations as the age label for each example. The length was recorded in centimeters and we show the length distribution for each year in Fig. 6. As is expected, older fish reach longer lengths on average but we see clear overlaps between years.

Ages of Atlantic cod, haddock and Greenland halibut ranged from 0 to 16, 1 to 15 and from 0 to 22 respectively. They were all read by a single reader. The resulting age distributions can be seen in Fig. A.1, Fig. A.2 and Fig. A.3 in the Appendix. Since there were few cases in the older age groups we clip the range of Atlantic cod to [1,10], haddock to [1, 8], and Greenland halibut to [4, 18]. We lump the older fish with the oldest age group in our clipped range. For Atlantic cod and haddock, no other features were available. For Greenland halibut, length measurements of the fish were recorded in centimeters, and the sex of the fish was also provided.

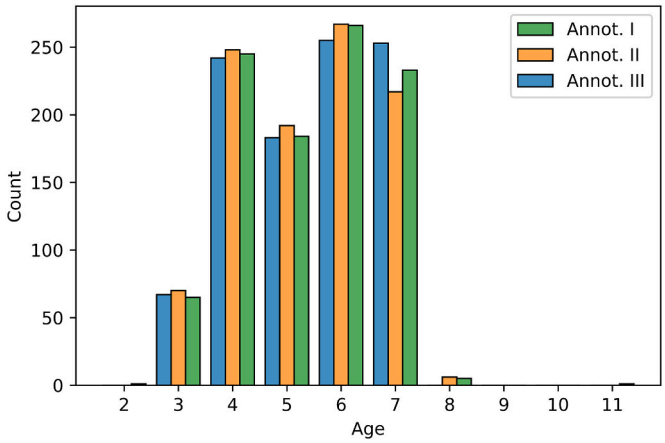


Fig. 4. Age distribution of the plaice otoliths for the three different annotators.

3.2. Model performance

We note that all experiments in Sections 3.2–3.5 were performed on plaice otoliths except for the evaluation of model performance on other species in Section 3.2.

For the CLIP-based approach, Fig. 7 shows the distribution of the difference in predicted age and annotated age for the CLIP models. Both models reached an age classification accuracy over 50%. The figure shows that most of the error predictions of the models, lay within ±1 year of the annotated age, and the errors were relatively evenly distributed between +1 and −1.

The classification report for the CLIP regression model is shown in Table 2. The precision, recall, and F1 score metrics indicated better performance for the age groups that had a higher number of observations. The model notably performed the worst on age group 3, which had the fewest numbers of samples.

The confusion matrix of the CLIP regression model is shown in Fig. 8. The matrix is normalized over the true age axis so the diagonal elements show the recall value for each age group. Looking at the figure, the highest values were obtained close to or on the diagonal. For true ages 4, 5, 6, and 7 years, the true positives have the highest recall. The model had the most trouble with the true age 3 years, which was the smallest age group. The model tended to overestimate the age of the younger fish (3–5 years old) while underestimating the age of seven year old fish.

Fig. 9 shows the difference in predicted age versus annotated age for the CLIP-regression model when the predicted age has been rounded to two decimals. Each bin in the figure has width 0.1. The figure shows that the difference in predicted age versus annotated age lies in a spread distribution curve with a narrow range, with no difference in predicted versus actual age, at the center of the curve. In Fig. B.1, in Appendix B, the difference in predicted versus true age for plaice can be seen when training the linear model on a training dataset of different sizes.

We further evaluated the CLIP regression method on the other species. Table 3 shows the comparison of average classification accuracy, the ±1 margin error accuracy, and the RMSE between the four different fish species datasets available in this study.

3.3. Comparison with other models

Fig. 10 shows the plaice test set accuracy for the CLIP models and the fine-tuned models. The CLIP regression model performs the best achieving 55.9% accuracy with RMSE of 0.70 years, and 97.05% accuracy when allowing a ±1 year margin of error like presented in Table 4.

The Inception models reached a considerably lower accuracy than the other models, and we found it overfitting the training data considerably faster compared to ResNet and ViT. Overall the CLIP models performed considerably better than the fine-tuned models. Comparing the CLIP regression model to the ViT fine-tuning we see a 5.6% increase in accuracy and a 0.19 decrease in the RMSE value.

3.4. Feature ablation

Table 5 shows the average classification accuracy (%) and standard deviation when different combinations of features available for plaice are used for the otolith age determination when using the linear and multiclass models. The linear model was able to obtain 55.90% accuracy when CLIP feature vectors, the quarter the fish was caught in, and the length of the fish, were used as input to the models. Compared to the multiclass model that obtained 54.15% accuracy for the same combination of input. Including the length of the fish in the input had a minor positive impact on the classification accuracy when CLIP features were present.

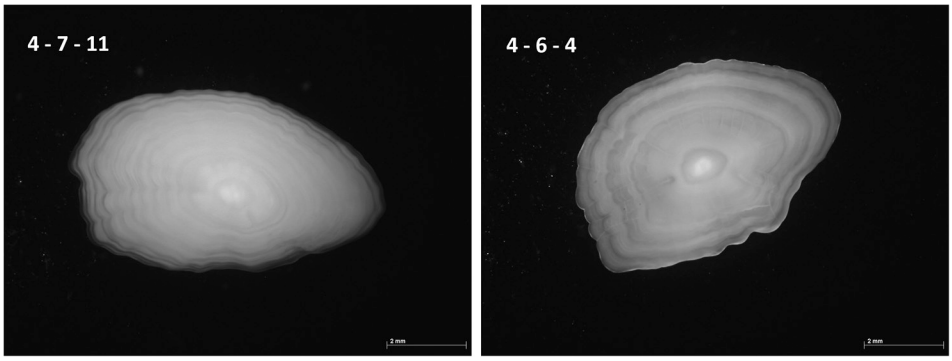


Fig. 5. Examples of annotator disagreement on plaice otoliths. The corresponding age annotations have been added to the upper left corner of the images.

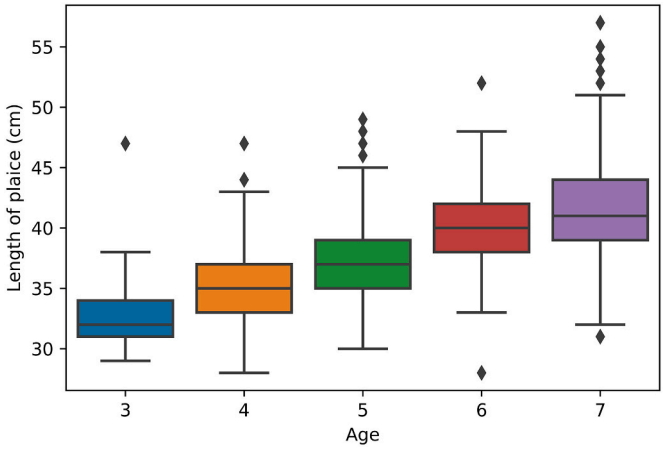


Fig. 6. Length distribution of plaice for each year of age. The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to the last datums within distance 1.5 times the interquartile range. Outliers are shown as individual points.

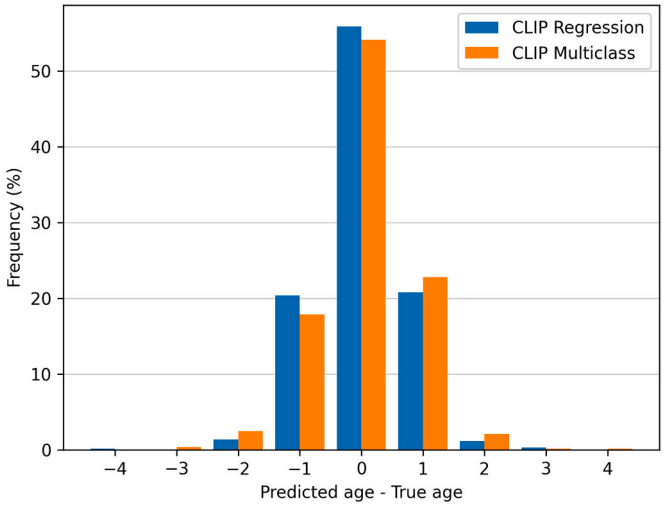


Fig. 7. CLIP regression vs. multiclass comparison. The distribution of the difference in predicted age versus true age is shown for the CLIP models using all data splits on plaice otoliths.

3.5. Few-shot performance

Fig. 11 compares the performance of the CLIP regression model and fine-tuning of the ViT when different dataset sizes were used for training. The classification accuracy of the CLIP regression model rapidly increased until the training dataset size reached 200 images. When the training dataset size increased further, the accuracy slowly increased. When the training dataset size reached 650 images, it appeared that the model was still learning as the classification accuracy was still increasing. When fine-tuning the ViT model the classification accuracy rapidly increased until the size of the training dataset reached 300 images. With increased training dataset size after 300, the accuracy did not grow as rapidly as before. Using 100 training samples the CLIP regression model surpassed the Inception models trained on a full training set. Figs. D.1–D.3, in Appendix D, show the learning curves for

Table 2
The test set classification report for the CLIP regression model on plaice otoliths. The precision, recall and F1 score metrics were obtained by computing the mean overall data splits.

Age	Precision	Recall	F1 Score	Support
3 years	0.49 ± 0.21	0.31 ± 0.17	0.38 ± 0.18	14
4 years	0.65 ± 0.07	0.58 ± 0.08	0.62 ± 0.06	50
5 years	0.44 ± 0.04	0.58 ± 0.10	0.50 ± 0.06	37
6 years	0.51 ± 0.04	0.64 ± 0.06	0.57 ± 0.05	52
7 years	0.72 ± 0.06	0.51 ± 0.08	0.60 ± 0.06	47
macro average	0.56 ± 0.05	0.52 ± 0.04	0.53 ± 0.05	200
weighted average	0.58 ± 0.03	0.56 ± 0.03	0.56 ± 0.04	200
accuracy			0.56 ± 0.03	200

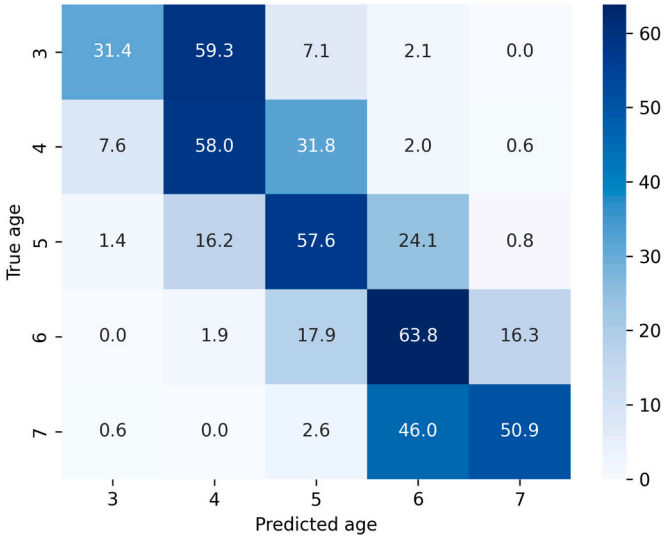


Fig. 8. Confusion matrix for the CLIP regression model. Results are shown for plaice normalized over the annotator age axis (shown as true age). Values are percentages (%).

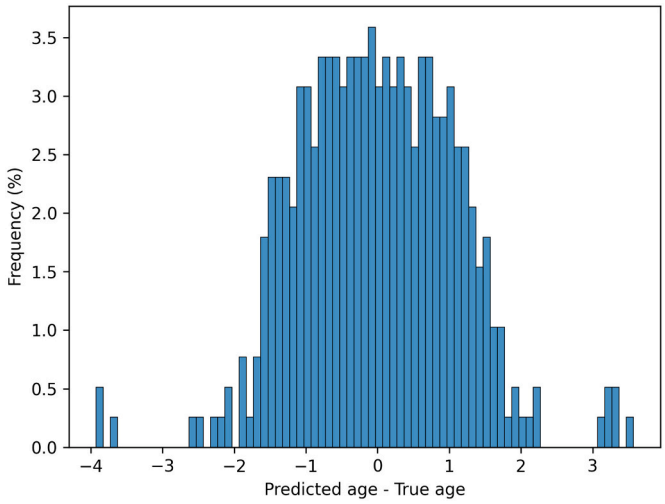


Fig. 9. Difference in predicted versus true age of plaice for the CLIP regression model (x-axis). The bin width is 0.1. Y-axis shows the frequency (%) of each bin.

Table 3
Average classification accuracy, accuracy with ±1 margin error, and RMSE for different fish species obtained by using the linear model. All numbers are presented with a standard deviation estimate.

Fish species	Accuracy (%)	±1 Accuracy (%)	RMSE
European plaice	55.90 ± 3.49	97.05 ± 0.75	0.70 ± 0.06
Atlantic cod	50.47 ± 2.37	94.10 ± 1.24	0.84 ± 0.04
Haddock	61.18 ± 1.70	96.53 ± 0.57	0.72 ± 0.03
Greenland halibut	29.76 ± 1.26	72.21 ± 1.29	1.42 ± 0.03

Atlantic cod, haddock and Greenland halibut using the CLIP regression model.

4. Discussion

In this study, we proposed a computer vision-based few-shot learning method for otolith age determination in European plaice, Atlantic cod, Greenland halibut, and haddock. Our method outperformed prior state-of-the-art approaches, achieving similar performance with up to three times less training data. One of the key advantages of our approach is its potential for transferability to other fish species, making it a general method that could be useful for otolith age determination in a variety of species.

For plaice otoliths, the few-shot approach with a CLIP regression model achieved the highest classification accuracy when comparing six

Table 4
Comparison of CLIP regression and multiclass approaches with fine-tuning methods for plaice age determination. All numbers are presented with a standard deviation estimate computed for a 10-fold cross-validation.

Model	Accuracy (%)	±1 Accuracy (%)	RMSE
CLIP Regression	55.90 ± 3.49	97.05 ± 0.75	0.70 ± 0.06
CLIP Multiclass	54.15 ± 2.06	94.80 ± 1.72	0.81 ± 0.05
ViT Fine-tuning	50.30 ± 3.06	93.10 ± 1.43	0.89 ± 0.04
ResNet Fine-tuning	49.40 ± 2.27	92.35 ± 1.75	0.88 ± 0.05
Deep Otolith Inception Fine-tuning	46.80 ± 3.62	90.20 ± 3.21	0.95 ± 0.05
Vanilla Inception Fine-tuning	45.15 ± 3.31	88.95 ± 1.94	1.00 ± 0.06

Table 5
Average classification accuracy and standard deviation estimate when using different combinations of plaice features available for otolith age determination. All numbers are presented with a standard deviation estimate.

Features	Regression	Multiclass
Length	35.20 ± 2.30	26.15 ± 0.34
Length + quarters	43.70 ± 3.35	37.65 ± 2.40
CLIP features	49.60 ± 3.53	50.25 ± 3.37
CLIP features + length	49.70 ± 3.58	50.20 ± 3.46
CLIP features + quarters	55.80 ± 3.97	54.15 ± 2.17
CLIP features + length + quarters	55.90 ± 3.49	54.15 ± 2.06

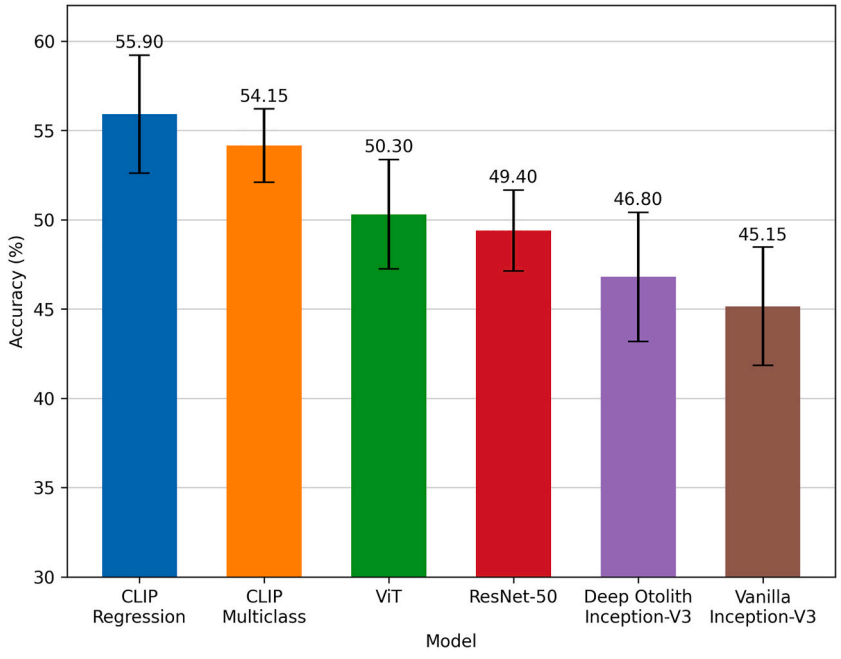


Fig. 10. Model accuracy comparison for plaice otoliths. Error bars show a standard deviation estimate for a 10-fold cross-validation.

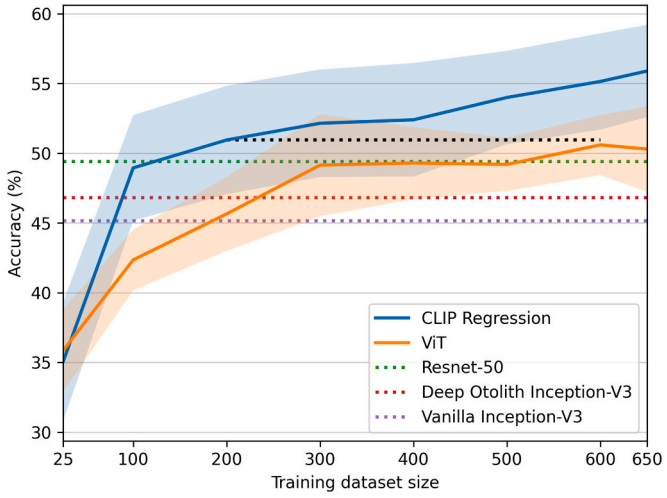


Fig. 11. Learning curve comparison of the CLIP regression model versus ViT fine-tuning for plaice otoliths. Using 200 samples for training the CLIP regression model reached the same accuracy as ViT trained on 600 samples as illustrated with a black dotted line. The figure also shows the average accuracy for the ResNet and Inception models using all training samples. The interesection with the blue curve demonstrates at which point the CLIP Regression reaches the same performance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different models. The CLIP regression model had a classification accuracy of 55.90% and 97.05% with a ±1 margin of error. Importantly, using CLIP vectors as features (along with quarters and fish length) for a linear model resulted in better performance than fine-tuning models like ResNet and Inception-V3 on the image input along with quarters and fish length. This difference is surprising, since fine-tuning deep networks like Inception-V3 is computationally much more expensive than training a linear model, and should, in theory, be a more flexible approach since all network parameters can be updated.

Including the length of the fish in the input improved the classification accuracy of our CLIP-based models on plaice otoliths by 0.1%, as seen in Table 5. Although older fish tend to be longer than younger ones, as seen in Fig. 6, there are significant overlaps between years. This result does not necessarily mean that fish length is not useful for age determination, as it may provide additional information beyond the signal from the otolith image and the quarter variable. It is also possible that the CLIP-based method is not able to make good use of the information provided by the fish length variable. An additional factor to consider is that plaice are sexually dimorphic, and using the age and gender variables together could potentially lead to improved performance.

The otoliths of Greenland halibut were found to be particularly difficult for our machine learning models to accurately determine the age of the fish. This challenge may be partly due to the fact that even human annotators struggle with this task (ICES, 2016). This difficulty is reflected in the results shown in Table 3, where the accuracy of the CLIP method is substantially lower for Greenland halibut otoliths but is in line with the expected ageing error for that species (and what has been observed in Ordoñez et al., 2022, and related studies). Albert et al. (2009) showed that previous methods for determining the age of Greenland halibut using otolith annuli counting had greatly underestimated the age of older fish. They also found that the average annual growth rate of adult halibut was just under 1 cm per year. They developed a new method that demonstrated a strong relationship between various measures of otolith size and age, after correcting for fish length. This suggests that otolith size may be a useful predictor of fish age and could potentially be learned by a machine-learning method.

Another feature that has been studied for estimating fish age is otolith weight, which increases as the fish grows. However, this approach may not be effective for older fish, as the changes in otolith weight may become smaller with age (Hanson and Stafford, 2017). It would be interesting to investigate whether incorporating otolith weight into our machine learning models could improve their performance.

In addition to adding more features to improve our machine learning models, we also studied the value of adding more training data. The learning curve in Fig. 11 shows that our models achieved relatively high accuracy even with a training dataset of only 100 images. The CLIP regression model had an accuracy of 48.95%, while the ViT fine-tuning model had an accuracy of 42.35%. The slope of the learning curve suggests that the CLIP regression model benefits more from additional training data than the ViT fine-tuning model in the range of hundreds to a few thousand examples. It is unclear, however, whether the fine-tuning approach could eventually outperform the CLIP regression model, or how many labeled examples would be needed for this to happen. It is also unclear at what point the accuracy of the models will reach a saturation point, and whether they will be able to achieve the 82% agreement of the annotators.

In our study, we used a reader's determination of a fish's age as a label to train our machine learning models. However, these labels were not validated through a particular method. Previous research has shown that age determination can be challenging. For example, a bomb radiocarbon validation study on haddock found that age was usually underestimated (Francis et al., 2010). In a mark-recapture study on Greenland halibut, the accuracy of age determination for four readers ranged from 68 to 89% (Albert, 2016). In our study on plaice, the interreader agreement was high (82%), but we did not have a reference point to confirm the true age of the otoliths. Etherton, 2015 showed in a mark-recapture experiment on plaice that the agreement of annotators with the predicted age based on the mark was around 25–33%, but by choosing the modal age of four annotators, the agreement reached 49.3%. Their results indicated that age was underestimated, but the reasons for this were not clear. In a markrecapture experiment on cod in the Baltic sea, researchers found that translucent zones in the otolith form during winter and are surrounded by opaque zones, supporting the practice of counting translucent zones to determine age (Krumme et al., 2020). However, the agreement of annotators with predicted age from mark-recapture experiments has been reported to be around 85% (ICES, 2019).

Given the classification results presented in this paper, one might ask how applicable they are. The automatic classification of fish ages has significant implications in the field of fisheries science, where the age structure of a fish population is critical to determining its productivity. The traditional method of age determination, which requires the expertise of skilled personnel, is both time-consuming and resource-intensive. The implementation of automatic classification has the potential to increase the efficiency and productivity of the age readers, who can now concentrate their efforts on improving the age-determination process.

While the current level of accuracy of automatic classification may be slightly lower than that of human readers, it has the advantage of processing a larger volume of otoliths in a shorter amount of time. This increased uncertainty can be addressed in difference equation models using separate likelihood functions for accounting for manual and automatic classifications. Combining the strengths of both methods has the potential to obtain a more accurate and comprehensive understanding of the age structure of a fish population.

Comparing the results of our study with previous research on otolith age determination can be difficult due to differences in species, age composition, dataset size, and annotation methods. These differences can largely explain why the accuracy of deep learning-based approaches to otolith age determination varies between studies. Previous studies often used different fish species, which can affect the difficulty of estimating the age of the fish based on their otoliths. In addition, the age composition of the datasets used in different studies can vary, which can affect the overall accuracy of the age estimates (it is usually easier to

estimate the age of younger fish). The way in which the otoliths are annotated can also influence the results, as different annotators may use different standards or criteria when labeling the data. Most previous studies on otolith age determination have used the Inception-V3 network, but our comparison indicates that using CLIP-based regression is a superior approach for plaice otoliths. We therefore encourage other researchers in the field to explore this computationally efficient method for their image classification tasks.

We can compare our results from the Greenland halibut model with other research using Greenland halibut otoliths. Martinsen et al. (2022) reports lower classification accuracy than we do. Our classification accuracy is similar to that reported in Moen et al. (2018), but our accuracy is slightly higher when allowing for a ±1 margin error. However, our average RMSE is lower than reported in both of these previous studies. It is important to acknowledge that these are all different datasets, even though they all include Greenland halibut otoliths. This may explain some of the differences in the results.

Based on our results, there is room for improvement in using computer vision techniques for otolith age determination. This will depend on an increased availability of annotated images from a wider age range and more species. Our analysis was limited by the available images, but in the future, the aim is to ensure that images of all otoliths are produced during routine age-reading. This has several benefits, including training new prediction models and improving between-reader validation, as otolith images can be shared with multiple locations. Additionally, Ordoñez et al., 2022 showed that trained models can also be shared across labs with an acceptable increase in uncertainty. While additional image standardization may be needed, this could allow expertise to be shared between labs.

Computer vision techniques could also allow for a more detailed exploration of age determination error and bias. Currently, this is mostly estimated based on multiple readings of the same otolith, and procedures are in place in most labs to minimize this error. When sufficiently trained, a model like the one we presented here may allow for inter-reader comparisons without the extra labor required to re-read the otoliths.

Overall, our paper demonstrates that the CLIP regression model with more samples can alleviate some of the manual labor involved in age-reading, making age readers more efficient.

Author contributions

All authors conceived the concept. B.Þ.E. and A.J. directed data collection. A.R.S., Þ.S. and H.E. conducted the initial data analysis and B.Þ.E. reviewed it. A.R.S., Þ.S., and H.E. wrote the main manuscript text and prepared figures. All authors reviewed the manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Data availability

The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

Acknowledgements

We would like to thank the Icelandic Centre for Research (Grant 2210656-1101) that funded Þ.S. and A.R.S. when working on this project in the summer of 2022.

Appendix A. Age distributions

The age distributions for Cod is shown in Fig. A.1, for Haddock in Fig. A.2, and for Greenland halibut in Fig. A.3.

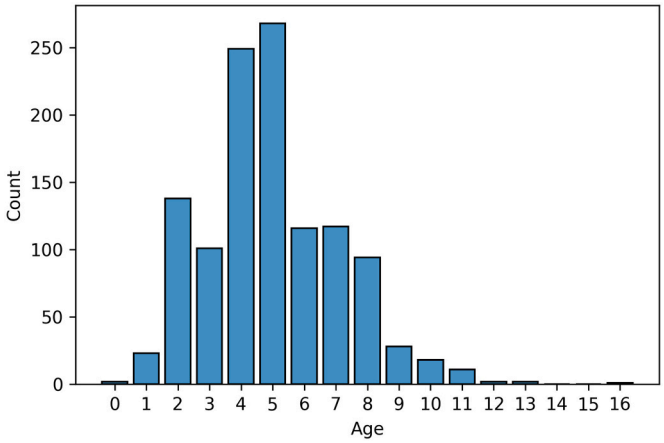


Fig. A.1. Age distribution of the Atlantic cod otoliths.

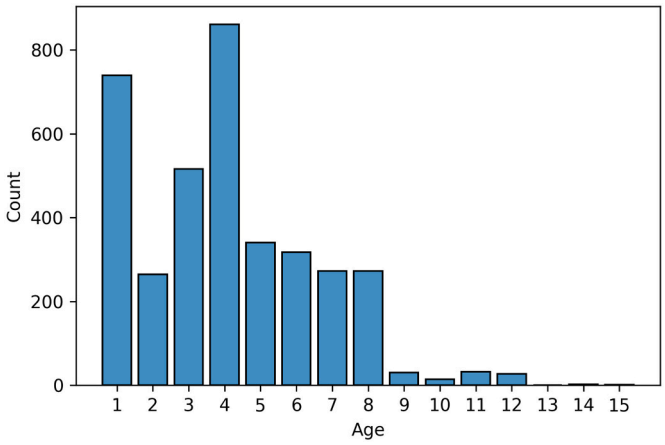


Fig. A.2. Age distribution of the haddock otoliths.

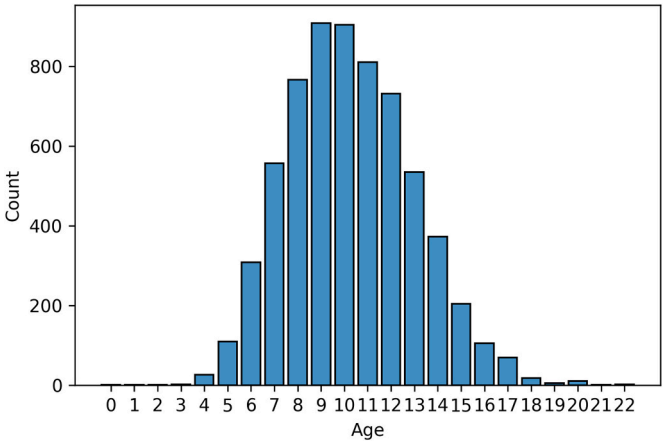
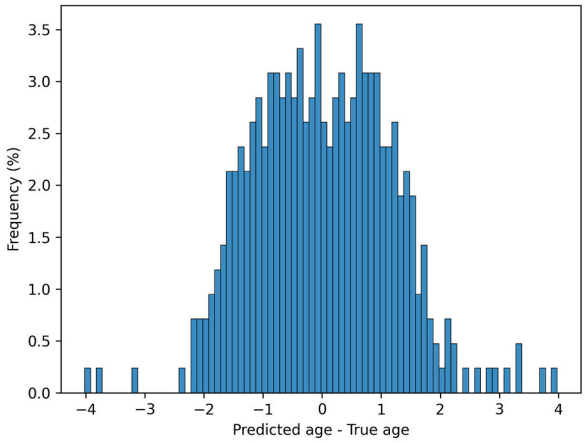
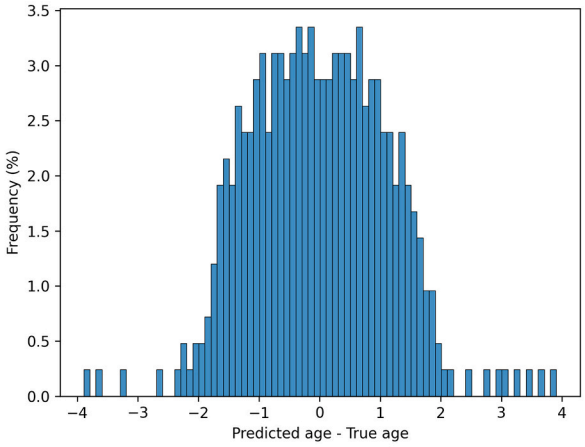


Fig. A.3. Age distribution of the Greenland halibut otoliths.

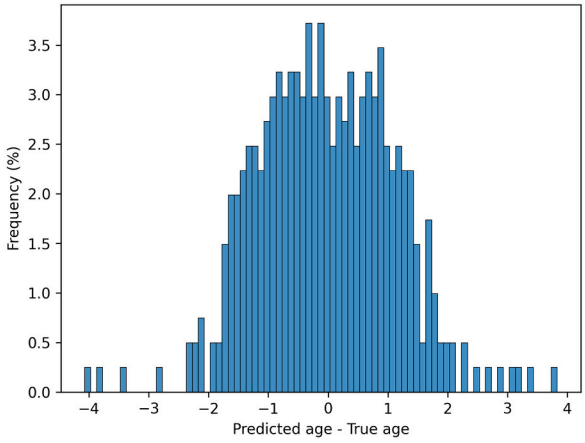
Appendix B. Difference in predicted versus true age for different training size



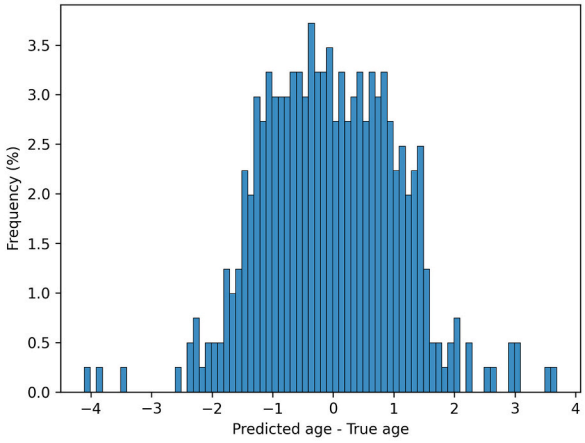
(a) 100 examples



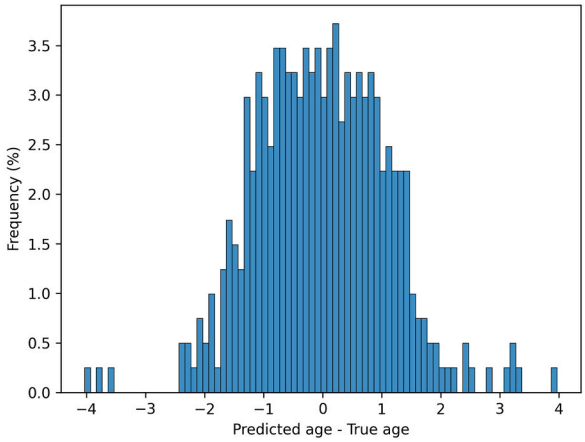
(b) 200 examples



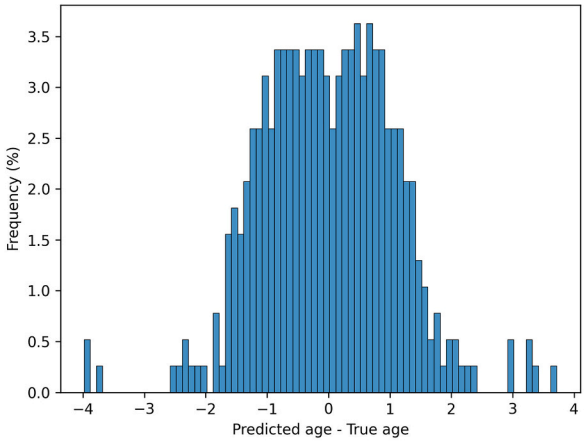
(c) 300 examples



(d) 400 examples



(e) 500 examples



(f) 600 examples

Fig. B.1. Difference in predicted versus true age for European plaice when training the linear model on a training dataset of different size. Training size is shown under each image.

Appendix C. Hyperparameter search

We performed hyperparameter search for our classification and regression model to identify the right level of regularization. For the linear regression model, we considered values of α that determine the L2 regularization to be in the range {0.1, 0.6, 1.1, …, 19.1, 19.6}. The resulting accuracy of the model is shown in Fig. C.1. We observe that the model shows stable performance for α values in the interval [5.0, 15.0]. For the

classification model, we considered values of C, the regularization parameter, to be in the range {0.001, 0.006, 0.011, …, 0.996, 1.001}. The resulting accuracy of the model are shown in Fig. C.2. When we are not using embedding features we used $\alpha = 0.1$ and $C = 0.1$, which were found through a hyperparameter sweep.

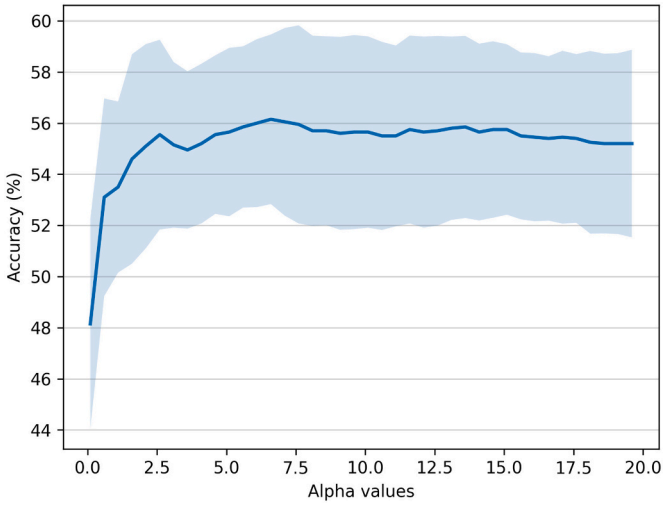


Fig. C.1. Accuracy of the linear regression model as a function of α, the L2 regularization parameter. Each experiment is repeated ten times and the envelopes represent a standard deviation estimate.

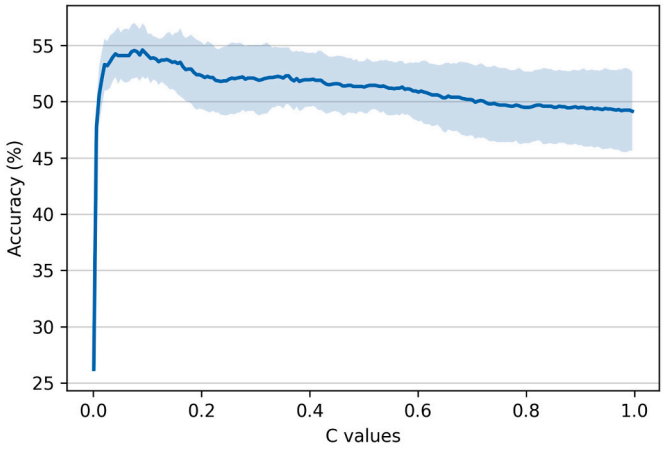


Fig. C.2. Accuracy of the classification model as a function of C, the regularization parameter. Each experiment is repeated ten times and the envelopes represent a standard deviation estimate.

Appendix D. Learning curves

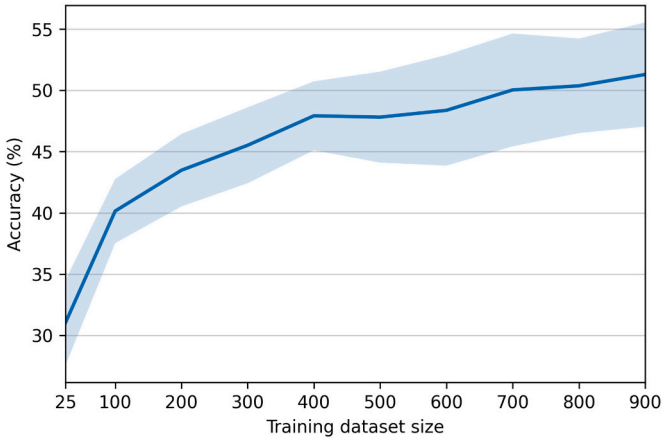


Fig. D.1. Accuracy learning curve for Atlantic cod using the CLIP regression model.

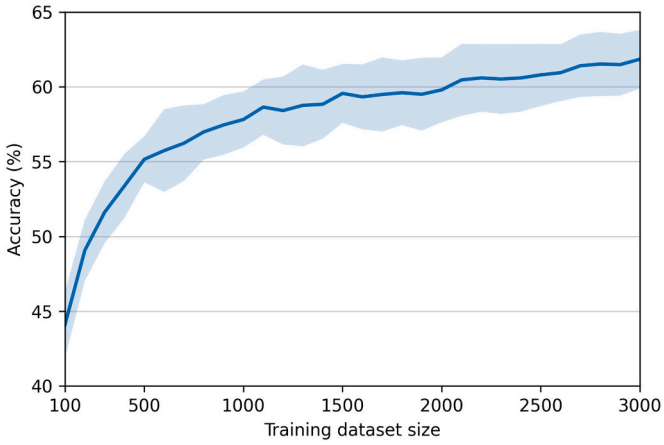


Fig. D.2. Accuracy learning curve for haddock using the CLIP regression model.

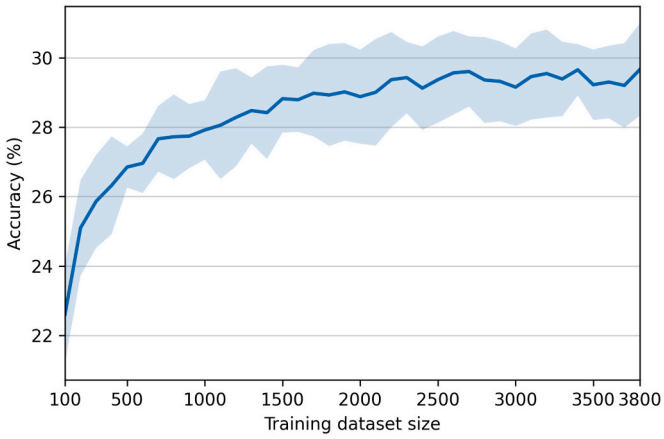


Fig. D.3. Accuracy learning curve for Greenland halibut using the CLIP regression model.

References

Albert, O.T., 2016. Growth and formation of annual zones in whole otoliths of Greenland halibut, a slow-growing deep-water fish. Mar. Freshw. Res. 67 (7), 937–942.

Albert, O.T., Kvalsund, M., Vollen, T., Salberg, A.-B., 2009. Towards accurate age determination of Greenland halibut. J. Northwest Atl. Fish. Sci. 40, 81–95. https://doi.org/10.2960/J.v40.m659.

Carbonara, P., Follesa, M.C., 2019. Handbook on age determination: A Mediterranean experience. In: General Fisheries Commission for the Mediterranean. Studies and Reviews, p. 98.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv: 1810.04805]. arXiv: 1810.04805 [cs]. Retrieved March 25, 2022, from. http://arxiv.org/abs/1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2023. February 26). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy.

Etherton, M., 2015. European plaice (*pleuronectes platessa*) and sole (*solea solea*) indirect age validation using otoliths from mark-recapture experiments from the north sea. Fish. Res. 170, 76–81.

Francis, R.C., Campana, S.E., Neil, H.L., 2010. Validation of fish ageing methods should involve bias estimation rather than hypothesis testing: a proposed approach for bomb radiocarbon validations. Can. J. Fish. Aquat. Sci. 67 (9), 1398–1408.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Grønkjær, P., 2016. Otoliths as individual indicators: a reappraisal of the link between fish physiology and otolith characteristics. Mar. Freshw. Res. 67 (7), 881–888.

Hanson, S.D., Stafford, C.P., 2017. Modeling otolith weight using fish age and length: applications to age determination. Trans. Am. Fish. Soc. 146 (4), 778–790.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition [arXiv:1512.03385 [cs]]. Retrieved January 26, 2022, from. http://arxiv.org/abs/1512.03385.

ICES, 2016. Report of the workshop on age reading of greenland halibut, 2 (wkargh2).

ICES, 2019. Benchmark Workshop on Baltic Cod stocks (WKBALTCOD2). https://doi.org/10.17895/ices.pub.4984.

Katayama, S., 2018. A description of four types of otolith opaque zone. Fish. Sci. 84 (5), 735–745.

Krumme, U., Stötera, S., McQueen, K., Pahlke, E., 2020. Age validation of age 0-3 wild cod gadus morhua in the western Baltic Sea through mark-recapture and tetracycline marking of otoliths. Mar. Ecol. Prog. Ser. 645, 141–158.

Martinsen, I., Harbitz, A., Bianchi, F.M., 2022. Age prediction by deep learning applied to Greenland halibut (*Reinhardtius hippoglossoides*) otolith images [Publisher: public library of science]. PLoS One 17 (11), e0277244. https://doi.org/10.1371/journal.pone.0277244.

Moen, E., Handegard, N.O., Allken, V., Albert, O.T., Harbitz, A., Malde, K., 2018. Automatic interpretation of otoliths using deep learning [Publisher: public library of science]. PLoS One 13 (12), e0204713. https://doi.org/10.1371/journal.pone.0204713.

Moore, B.R., Maclaren, J., Peat, C., Anjomrouz, M., Horn, P., Hoyle, S.D., 2019. Feasibility of Automating Otolith Ageing Using CT Scanning and Machine Learning [Publisher: Fisheries New Zealand]. https://doi.org/10.13140/RG.2.2.29670.16960.

Ordoñez, A., Eikvil, L., Salberg, A.-B., Harbitz, A., Elvarsson, B.Þ., 2022. Automatic fish age determination across different otolith image labs using domain adaptation [number: 2 Publisher: multidisciplinary digital publishing institute]. Fishes 7 (2), 71. https://doi.org/10.3390/fishes7020071.

Panfili, J., de Pontual, H., Troadec, H., Wrigh, P.J., 2002. Manual of fish sclerochronology. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

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. https://doi.org/10.1016/j.fishres.2021.106033.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models from Natural Language Supervision.

Shaffi, N., Hajamohideen, F., 2021. Few-shot learning for Tamil handwritten character recognition using deep Siamese convolutional neural network. In: Mahmud, M., Kaiser, M.S., Kasabov, N., Iftekharuddin, K., Zhong, N. (Eds.), Applied Intelligence and Informatics. Springer International Publishing, pp. 204–215. https://doi.org/10.1007/978-3-030-82269-9_16.

Sólmundsson, J., Kristinsson, K., Steinarsson, B., Jonsson, E., Karlsson, H., Björnsson, H., Palsson, J., Bogason, V., Sigurdsson, T., Hjörleifsson, E., 2010. Manuals for the Icelandic Bottom Trawl Surveys in Spring and Autumn. Marine Research Institute, Reykjavík, Ice-land.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015a. Going deeper with convolutions. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015b. Rethinking the Inception Architecture for Computer Vision [arXiv:1512.00567 [cs]]. https://doi.org/10.48550/arXiv.1512.00567.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30.

Wang, C.-H., Walther, B.D., Gillanders, B.M., 2019. Introduction to the 6th international otolith symposium. Mar. Freshw. Res. 70 (12), i–iii.

Xu, Z., Zhu, L., Yang, Y., 2016. Few-Shot Object Recognition from Machine-Labeled Web Images [arXiv: 1612.06152]. arXiv:1612.06152 [cs]. Retrieved January 6, 2022, from. http://arxiv.org/abs/1612.06152.