Development of a Predictive Model for Estimating Fish Age from Otolith Images in the Southern Gulf of St. Lawrence

WatSpeed Machine Learning Course - Final Report (Group 11)

Chen, Xin[1], Fishman, David[2], Shan, Xiaojin[3], Thaker, Rudra[4], Thankappan, Shinoj[5], and Wu, Xuan[6]

Emails: [1]x75chen@uwaterloo.ca, [2]dfishman@uwaterloo.ca, [3]x23shan@uwaterloo.ca, [4]r2thaker@uwaterloo.ca, [5]sthankap@uwaterloo.ca, [5]x34wu@uwaterloo.ca

December 2024

Contents

1 Objectives

Monitoring fish stocks is a critical component of sustainable fisheries management in the Southern Gulf of St. Lawrence. One key aspect of this work is understanding the population dynamics of various fish stocks, which involves accurate age determination. Age data are essential for modeling growth patterns, understanding reproduction, and assessing the health and sustainability of fish populations.

1.1 Goal of the analysis

The primary objective of this project is to develop and implement a machine learning-based predictive model capable of estimating the age of fish from otolith images. This model will automate the process of age determination, reduce the potential for human error, and provide quicker assessments for large datasets.

Specifically, the objectives of the project include:

- Data Collection and Preparation: Compile an archive of otolith images along with corresponding fish age, length, and weight data.
- Model Development: Create a predictive model using machine learning techniques to automatically identify annuli in otolith images and predict the age of the fish.
- Model Validation: Validate the model's accuracy using a separate set of otolith images and corresponding age data.

The training dataset will consist of fish otolith images originating from two Atlantic Canadian species of economic and ecological importance: American plaice (*Hippoglossoides platessoides*) and Atlantic herring (*Clupea harengus*).

1.2 Rationale behind the analysis

Fish age is commonly determined by examining biological materials such as otoliths (inner ear bones) and scales. These materials exhibit growth rings, or "annuli" (Figure 1 and Figure 2), which can be counted similarly to tree rings. Each ring represents a period of growth, typically corresponding to one year in the life of the fish. However, manually counting these rings can be time-consuming, subjective, and prone to human error. Organizations

that age fish are typically limited by the amount of time it takes the team of human experts to read the otolith samples.

The development of an automated system for predicting fish age based on otolith images would significantly improve the accuracy and speed of age estimation. At a minimum, having a machine learning (ML) model to complement the work of human agers will provide an aspect of quality control to the aging process. At best, these models might actually help speed up the processing time required for aging fish.

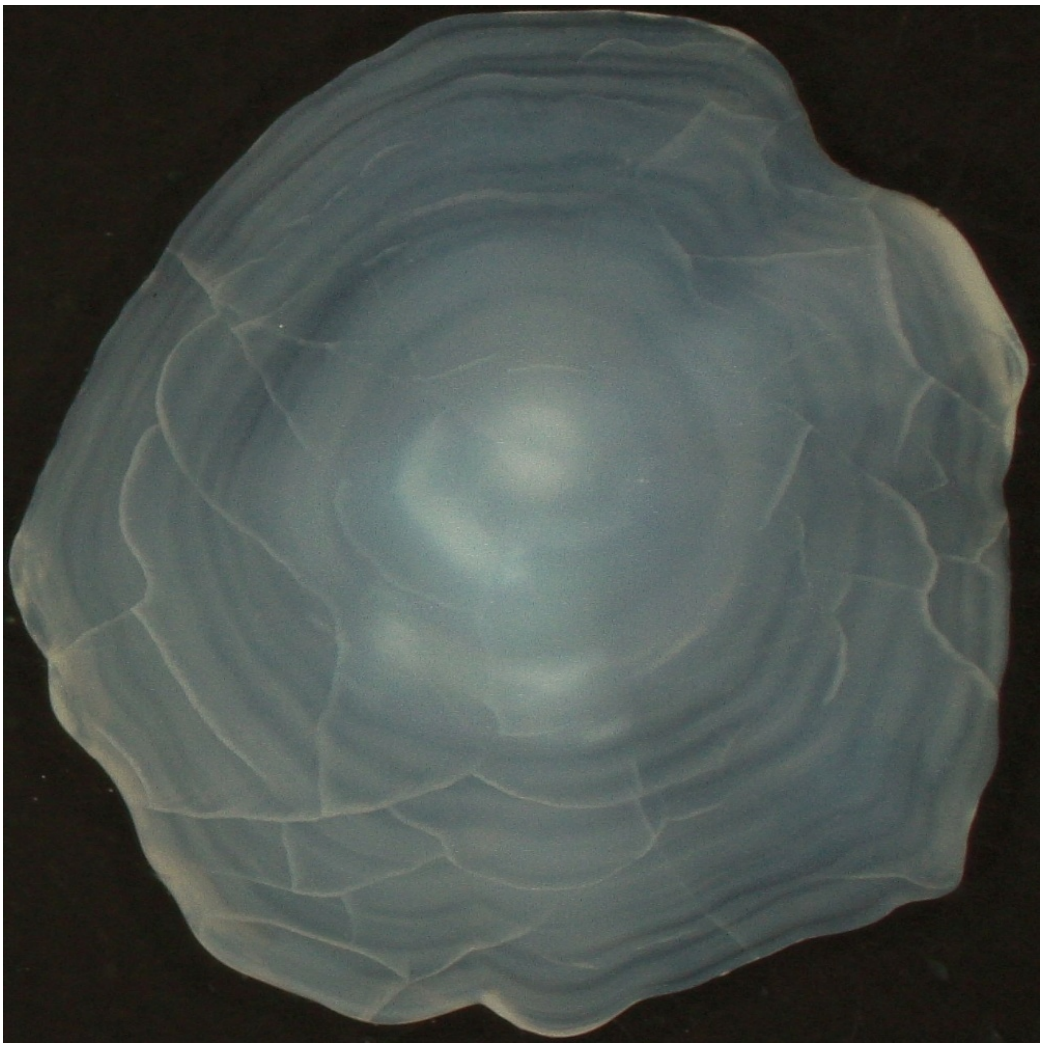

Figure 1: An example of an otolith image taken from an American Plaice.

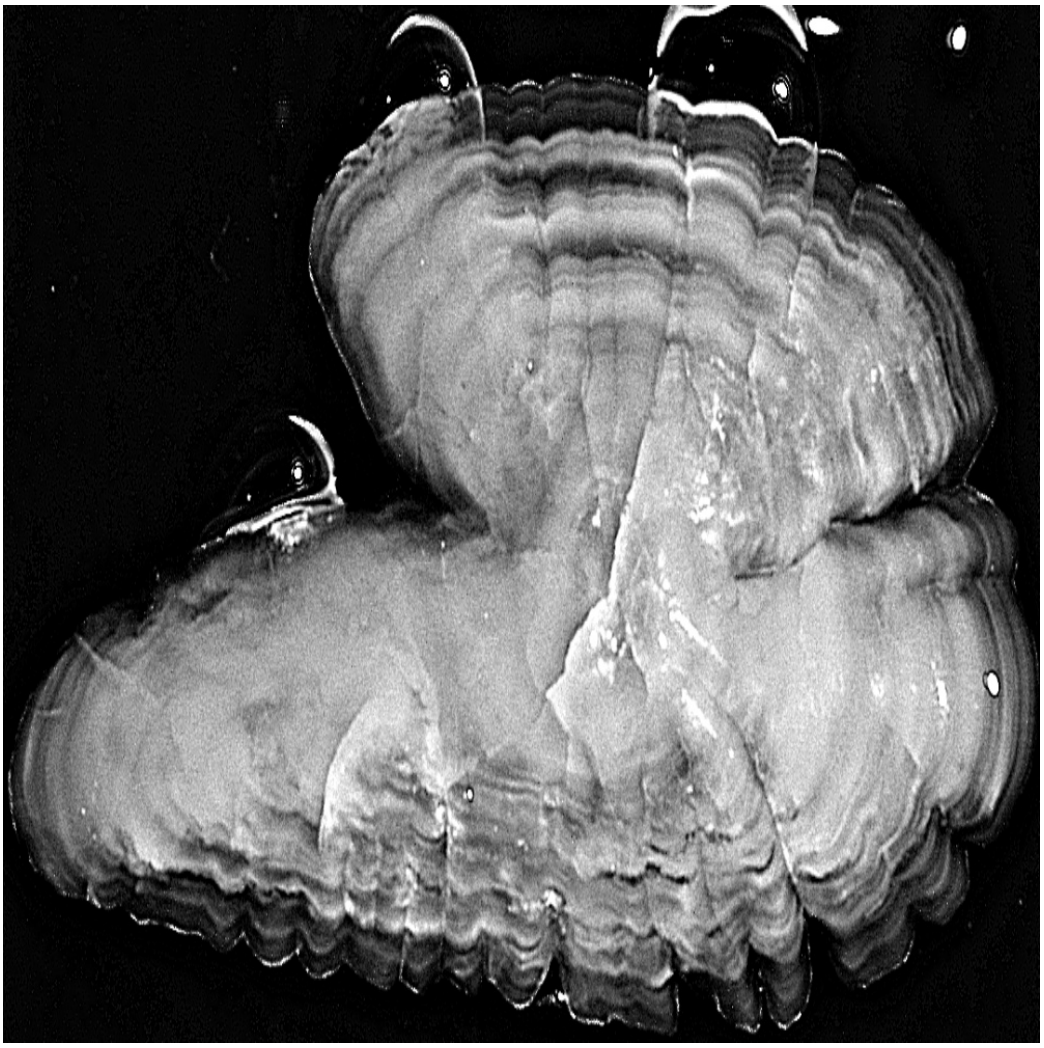

Figure 2: An example of an otolith image taken from an Atlantic Herring.

2 Data Preparation

2.1 What was your data source?

The data used for this analysis originate from archives of images taken of otoliths from two Atlantic Canadian species of fish: Atlantic Herring and American Plaice. While these archives are unpublished, some of the details about how the otoliths were collected, preserved and imaged are available on the Government of Canada's Open Data Portal [3]. The fish specimens themselves were collected from various Canadian government stock assessment surveys (e.g., [1] and [2]) as well as from commercial port sampling programs.

2.2 Data quality and procurement?

The quality of the data was very high. All the specimens had unique universal identifiers (UUIDs) which corresponded to the image file names. None of the images used were orphaned, i.e., without corresponding metadata.

The dataset contained ages that were evaluated by human experts and which are actively used in annual population models for the two species in question. Furthermore, since each fish specimen has two otoliths, our target dataset was effectively doubled since for every fish whose age is known, there were two corresponding images.

The data was procured by a team member who has a role at Fisheries and Oceans Canada in maintaining and archiving the dataset.

2.3 What tools or code did you need to use to prepare it for analysis?

Here is a list of the python tools used for processing, data preparation and model construction:

- Jupyter Notebooks: for the development and presentations of our analysis
- OpenCV: for image processing and manipulation.
- CNN (TensorFlow / Keras): for deep learning and automatic feature extraction from images.
- XGBoost: for gradient-boosting regression modeling.

- Scikit-learn: for machine learning models, data preprocessing, and evaluation
- Pandas: for handling and manipulating tabular data
- Matplotlib / Seaborn: for visualizing data and model performance.

2.3.1 Fish Specimen Data

Information about the fish specimen was stored in a CSV file which contained metadata for each otolith, including the UUID, age, length, weight, and other categorical features such as gender and species. The categorical variables were already converted into dummy variables. For example, at the time of collection, a specimen's sex is identified as either 'male', 'female' or 'unknown'. In the dataset, these categories were represented by three mutually exclusive columns: 'is_male', 'is_female' and 'is_unknown'. The species were represented by two mutually exclusive columns: 'is_plaice' and 'is_herring'. As noted above, the metadata is linked to the corresponding otolith image based on the UUID. The CSV file was simply read into a pandas dataframe. In Table 1 a detailed explanation of each column in the dataset is provided.

2.3.2 Otolith Images

The dataset contained a total of 15,089 images of otoliths originating from the 7828 specimens. Therefore, on average, each specimen had 1.9 corresponding otoliths. The otolith image was identified by a UUID prefix in its filename. The original resolution of the images were 1000 x 1000 pixels and were in RGB. To make the image processing more manageable for this project, they were resized to 64 x 64 pixels and to a greyscale image. The image was subsequently flattened from a 1D array to a 2D array. Similarly, for the purposes of shortening the processing time of the scripts, only one randomly selected image from each specimen was used in the analysis. Finally, the images were read into a 7,828 x 4097 pandas dataframe, where the rows were the image instances and the columns were the pixels plus one extra one for the UUID which was parsed from the filename. This was then merged with the metadata dataframe referenced above using a left join on the 'uuid' column. After the images and metadata were merged, the numerical features of the resulting matrix were normalized. Specifically, we used the StandardScaler to normalize the numerical features (e.g, length, weight and pixel columns) meaning those columns will have zero mean and unit variance. This step is crucial for machine learning models to perform optimally. Finally, the

dataset was split into training and testing portions using the 'train_test_split' according to a ratio of 4:1, respectively. This resulted in 6262 instances for model training and the remaining 1566 instances for model testing.

2.4 What challenges did you face?

One of the main significant challenges encountered during this analysis was the large size of the dataset. With over 15,000 images totaling nearly 4 GB, managing and processing the data was difficult. The sheer volume of data made it difficult to disseminate and significantly slowed down the model training process, making it clear that resizing the images was necessary. As a result, we had to resize the images to very small sizes to achieve reasonable computation times. However, smaller images likely led to the loss of important details, which could potentially decrease accuracy. Another challenge involved managing multiple images per UUID. Since each specimen typically had more than one image, it was necessary to determine the most effective method for utilizing them.

Three approaches were explored:

1. **Random Selection:** Selecting a single image at random per UUID.
2. **Image Concatenation:** Merging multiple images into a single, larger input.
3. **Image Averaging:** Averaging the images to generate a single representative image.

Each approach presented its own limitations. Random selection could overlook valuable information, concatenation introduced additional complexity, and averaging risked the loss of critical features. Balancing accuracy with efficiency in choosing the optimal method proved to be a time-consuming and challenging task.

Table 1: Detailed explanation of the column in the specimen metadata dataset

Name	Type	Description	Purpose

uuid	String (or unique identifier)	A unique identifier for each data record. This is useful for tracking and referencing individual fish samples in the dataset.	Helps ensure each record is uniquely identifiable. It would not be used directly in the model for predictions, but it's essential for managing the dataset.
fish_id	String	A unique identifier for each individual fish.	Individual fish records. It might be useful for aggregation or analysis purposes but would generally not be used in the model itself.
age	Integer	The known age of the fish, which will be used as the target variable for the model.	The primary target variable to predict based on the otolith image features. The model will learn to predict this based on otolith image analysis and possibly other features.
length	Float	The length of the fish, which is often correlated with age and could be useful as a feature for the model.	Features like length and weight can help the model better estimate the age since they provide additional biological context for each fish.
weight	Float	The weight of the fish, which, like length, may correlate with age.	Another potentially important feature for age estimation, as larger or heavier fish may be older.

month	Integer	the specimen was collected (from the wild) normalized to 1	A potential feature for the model to understand seasonal variations in growth or age estimation.
is_male	Boolean (0 or 1)	A binary value indicating whether the fish is male (1) or not (0).	Gender could influence growth rates and age estimation, so this feature might improve model accuracy.
is_female	Boolean (0 or 1)	A binary value indicating whether the fish is female (1) or not (0).	Like is_male, gender could affect age estimation, especially in species where males and females have different growth patterns.
is_unknown	Boolean (0 or 1)	A binary value indicating unknown (1) or known (0).	Gender information is unavailable for these instances.
is_plaice	Boolean (0 or 1)	A binary value indicating whether the fish is an American plaice (1) or not (0).	Species-specific characteristics could influence the model, as growth patterns may differ between species. This feature will help the model differentiate between the two species.
is_herring	Boolean (0 or 1)	A binary value indicating whether the fish is an Atlantic herring (1) or not (0).	Similar to is_plaice, this feature will help the model differentiate between the two species (American plaice and Atlantic herring), as their growth rates, lifespan, and annuli in otoliths may differ.

3 Analysis

3.1 Methodology

First, we aimed to build a simple predictive model that integrates the image data with the metadata measurements, with the goal of outperforming the metadata alone. Comparison of SVM regressor (SVR), Elastic Net, Decision Tree, Random Forest, XGBoost, Voting, and Stacking (combining SVR, ElasticNet, Decision Tree, Random Forest, and XGBoost with the secondary model as indicated in the tables below) models were performed on the datasets. A randomized grid search was performed separately for each model to determine the optimal hyperparameters before comparing the models. Metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2) were used as key evaluators.

Next, we built models on the integrated metadata and image data aiming for better performance but applying a PCA transformation on the image data, preserving 95% of the variance. As a result, the dimensionality of the image data was reduced from 4,096 pixels to 340 dimensions

Finally, we explored a Convolutional Neural Network (CNN) model to further enhance our predictive performance. The CNN model is designed to automatically learn relevant features from the otolith images, which is particularly useful in identifying the age-related annuli in the images. The CNN architecture used in this analysis consists of several convolutional and pooling layers, followed by fully connected layers that allow for regression of the fish age.

3.1.1 CNN Model Architecture

The CNN architecture used in this analysis includes:

- **Convolutional Layers:** These layers apply filters to extract features from the images. We started with 32 filters in the initial layers, progressively increasing to 128 filters in the deeper layers.
- **Max-Pooling Layers:** These layers downsample the feature maps, reducing the computational load and aiding in generalization.
- **Flatten and Dense Layers:** The feature maps are flattened and passed through fully connected layers to predict the fish age.
- **Dropout Layer:** A dropout layer was incorporated to mitigate overfitting during training by randomly disabling a fraction of neurons.

The model was trained using the Adam optimizer and Mean Squared Error (MSE) as the loss function. We evaluated the model's performance using key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) on both the validation and test datasets.

3.2 Results

In Table 2, the results of the simple models are displayed. The stacking (gradient boost) model is the top performer for metadata, with the lowest MSE (1.385594), the lowest RMSE (1. 177113), and the highest R^2 (0.848093). Other ensemble models, such as Stacking (linear regression), Random Forest, XGBoost, and Voting, performs similarly well in terms of accuracy, with MSE below 1.5, RMSE below 1.25, and an R^2 above 0.83, slightly lower than the Stacking (gradient boost) model. Single models, such as SVR, Elastic Net, and Decision Tree lag behind the ensemble methods in model performance. Among them, the linear regression model Elastic Net shows drastically lower scores in evaluation metrics.

In Table **??**, we see the results from the PCA transformed data. Stacking (Linear regression) model is the top performer for metadata, with the lowest MSE (1. 514197), the lowest RMSE (1. 230527), and the highest R^2 (0. 833993). Similar to the models built on the metadata alone, we observed that the ensemble models outperformed the single models. Additionally, integrating image data improved the performance of some single models, particularly for the SVR and ElasticNet models. Incorporating image data did not significantly improve the performance of the ensemble models as we had initially expected.

Finally, the results of the CNN test and validation set are displayed in Table **??**. The CNN model achieved an R^2 of 0.73 on the test set, which is a promising result given the complexity of the task. The model demonstrated a MSE of 2.64 and RMSE of 1.62, indicating room for improvement, but also showing that it performs better than simple models based solely on metadata. The model showed better performance on the test set compared to the validation set, where the R^2 dropped to 0.55. Further tuning of the model, such as experimenting with more complex architectures or adjusting hyperparameters, may improve performance on both the validation and test sets.

Table 2: Performance summary of model built on metadata

Model	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-squared (R^2)
SVR	1.984	269 1.40864	1 0.782458
ElasticNet	2.104611	1.450728	0.769264
Decision Tree	1.579715	1.256867	0.82681
Random Forest	1.39943	1.182975	0.846576
XGBoost	1.484482	1.218393	0.837251
Voting	1.496319	1.223241	0.835953
Stacking (Linear regression)	1.391861	1.179772	0.847405
Stacking (Gradient boost)	1.385594	1.177113	0.848093

4 Conclusions

4.1 Was the model useful?

Lorem ipsum dolor sit amet, consectetur adipisicing elit. A, accusantium cum dolor eveniet facere impedit laborum non recusandae sit velit. Doloribus ducimus est exercitationem libero maxime nostrum omnis quos temporibus.

4.2 What did you learn about your data set?

Lorem ipsum dolor sit amet, consectetur adipisicing elit. A, accusantium cum dolor eveniet facere impedit laborum non recusandae sit velit. Doloribus ducimus est exercitationem libero maxime nostrum omnis quos temporibus.

References

[1] Government of Canada. Nafo division 4t groundfish research vessel trawl survey (september survey) dataset. https://open.canada.ca/data/en/dataset/1989de32-bc5d-c696-879c-54d422438e64, 2024. Accessed on 2024-04-13.

[2] Government of Canada. Northumberland strait multi-species trawl survey data. https://open.canada.ca/data/en/dataset/6d61f7b4-39eb-c8a9-a71c-b65d2eca8660, 2024. Accessed on 2024-12-08.

[3] Government of Canada. Otolith collection of american plaice in the southern gulf of st. lawrence. https://open.canada.ca/data/en/dataset/c1c739ee-a0ae-0e14-9c11-8568b75169b9, 2024. Accessed on 2024-12-08.