Automated Fish Size Measurement System for Long-Term Growth Studies in the Azores



Brenda Karoline Cruz Nogueira
Master in Data Science
Department of Computer Science
Faculty of Science of the University of Porto
2024

Automated Fish Size Measurement System for Long-Term Growth Studies in the Azores

Brenda Karoline Cruz Nogueira

Dissertation carried out as part of the degree in Master in Data Science

Department of Computer Science

2024

Advisor
Rita Paula Almeida Ribeiro, Auxiliar Professor, Faculty of Sciences, University of Porto

Co-advisors
Nuno Moniz, Associate Research Professor, Lucy Family Institute for Data & Society, University of Notre Dame

Gui Manuel Menezes, Director, Institute of Marine Sciences – OKEANOS, University of Azores

Abstract

Fishery analysis is essential for ensuring the long-term sustainability of marine species and supporting the livelihoods of millions who depend on fishing for food and income. Analyzing commercial catches is crucial for fishery stock assessment and management, as it provides estimates of fishing effort and its effects on marine ecosystems. Two main factors are critical in this analysis: first, the fishing gear, or metier, which significantly impacts marine habitats by targeting specific species and fish sizes. However, metier information is often incomplete in datasets. Second, fish size data is vital for studying the potential impacts of fisheries on marine ecosystems, but manually annotating daily landings is nearly impossible. This paper examines data from the Azores, including fishing records from 2010 to 2017 with limited metier information, and recent fish images and measurements from 2022 to 2024. Our study has two main objectives: first, to reconstruct incomplete datasets by applying imbalanced domain and machine learning techniques to link metier-related information with fish landings; second, to explore fish length prediction using computer vision methods for real-time estimation. We demonstrate that inference in metier classes can address similarity in fishery along with misclassification patterns in the dataset, while key image information can enhance predictions, aiding in the creation of automated systems essential for future fish population assessments, management, and conservation.

Acknowledgments

This thesis received national funds through the FCT – Foundation for Science and Technology, I.P., under the project UIDB/05634/2020 and UIDP/05634/2020 and through the Regional Government of the Azores through the project M1.1.A/FUNC.UI&D/003/2021-2024. We also acknowledge the LOTAÇOR, S.A., a regional public company responsible for the fish auctions and fisheries landings statistics, and João Santos (OKEANOS - UAc), which manages the LOTAÇOR/OKEANOS-UAC fishing landings database.

I would like to express my deepest gratitude to the people and entities who supported me throughout this project. To Professor Nuno Moniz for his invaluable support, guidance, trust, and encouragement over these years, and to Professor Rita Ribeiro, who trusted me as a student to undertake this work. To Professor Gui Menezes, who generously shared the data, time, and expertise crucial to completing this work. To my family, who, despite the distance, have been an inexhaustible source of strength and inspiration. I am deeply grateful to my mother, Elisangela Barros Da Cruz, and sister, Jade Nogueira, for believing in me and supporting me unconditionally. To my life companion and husband, Luis Oliveira, thank you for your unwavering love, encouragement, and support throughout these years; this work is also your achievement. To my grandma and aunts for all their support over my life. To my dog, who provided me with much-needed emotional support. To Gustavo Bongiovi, for your invaluable assistance with the ideas for this work and text development. To Veronica Bertolini, Bruna Meirelles, and all my friends, your friendship has been essential in keeping me motivated and balanced throughout this journey.

Contents

Abstract i

Acknowledgments iii

Contents vi

List of Tables vii

List of Figures ix

Acronyms xi

1 Introduction 1

2 Related Work 5

2.1 Data Reconstruction 5

2.2 Fish size prediction 6

2.2.1 Deep Learning 7

2.2.2 Fish Segmentation 8

2.2.3 Length Estimation 9

3 Data Presentation and Analysis 11

3.1 Datasets 11

3.1.1 Daily landings dataset 11

3.1.2 Fish Images 20

4 Methodology **23**

4.1 Data Reconstruction 23

4.1.1 Methods 24

4.2 Automated Fish Size Measurement 27

4.2.1 Pixel Length Distance 27

4.2.2 Research Questions 28

4.2.3 Methods 29

5 Results and Analysis **33**

5.1 Results Data Reconstruction 33

5.2 Discussion 36

5.3 Results Length Prediction 38

5.3.1 Segmentation 38

5.3.2 Feature Correlation 39

5.3.3 Regression Models Hyperparameters Search 40

5.3.4 Result on Test Set 40

6 Conclusion **43**

6.1 Limitations 44

6.2 Future Work 44

A Resampling Parameters in Data Reconstruction **45**

B Parameters Tested in Data Reconstruction Models **47**

C Hyperparameters for Length Prediction Algorithms **49**

Bibliography **51**

List of Tables

2.1 References resume for data reconstruction 6

2.2 References resume for fish length prediction using images 10

3.1 Features description 12

3.2 Metier description, the total weight of fish caught and the average weight for each classification, in kilograms (Kg) 13

3.3 Major fish classifications, total number of landings, total fish weight and 13

3.4 Species annotations 21

5.1 Accuracy of each metier class per method on the data set with new features . . . 34

5.2 Three Iconic Species of each metier 36

5.3 Confusion Matrix generate with ensemble strategy 37

5.4 Bounding Box and Segmentation Results using Yolo 38

5.5 Features Correlation with Total Fish Length 39

5.6 Best results for hyperparameter search 41

5.7 MAE Results of Models in Test Set 41

A.1 Best resampling strategies 45

B.1 Best Parameters For Random Forest and Boosting for each window strategy . . . 48

List of Figures

1.1 Summarization of the steps of the work. 3

2.1 Pipelined Architecture of Convolutional Neural Network (**source:** Sinha et al. [60]) 7

2.2 A sample annotated image illustrating the difference between image-level annotations, object-level annotations, and segmentations at the class/semantic- or instance-level. (**source:** Tedrake [64]) 8

3.1 Azores harbors and the respective number of landings 12

3.2 Total weight of each fish classification over the years 14

3.3 Distribution of major fish groups over the years per metier. 15

3.4 Distribution of major fish groups over the years per metier: small fish classes . . 15

3.5 Time Series of Islands, Metiers and Classifications 16

3.6 Networks over the years. 19

3.7 Example of a sample in the FishMetric's system. 21

3.8 Distribution of length for each specie. 21

4.1 26

4.2 Visualization of Fish Length Measurement in Pixels. 28

5.1 Machine learning methods and time windows 34

5.2 General Segmentation Model. 42

5.3 Specific Segmentation Model. 42

Acronyms

AP Average Precision

CPUE Catch per Unit Effort

CNN Convolutional Neural Networks

DNN Deep Neural Networks

DTW Dynamic Time Warping

EEZ Exclusive Economic Zone

FAO Food and Agriculture Organization of the United Nations

ICES International Council for the Exploration of the Sea

MAE Mean Absolute Error

mAP Mean Average Precision

ML Machine Learning

Mask R-CNN Mask Region-based Convolutional Neural Network

R-CNN Region-based Convolutional Neural Network

SAM Segment Anything Model

TACs Total Allowable Annual Catches

YOLACT You Only Look At Coefficients

YOLO You only look once

Chapter 1

Introduction

This dissertation leverages domain knowledge and machine learning tools to explore methods for long-term growth studies in the Azores Islands. Using data and the system provided by the Institute of Marine Sciences-OKEANOS, this dissertation has two main objectives: first, to develop techniques for data reconstruction to address gaps of information about fisheries gears in fisheries datasets, and second, to create automated methods for fish size measurement. The aim is to enhance the quality of fisheries databases, ultimately contributing to better-informed advice and sustainable management of wild fish populations over the years.

The global demand for seafood has increased substantially, with an estimated 223.2 million metric tons of fish production worldwide in 2022, of which 165 million metric tons were used for human consumption [22]. However, this high demand has also had an unprecedented impact on aquatic ecosystems and is believed to have caused a reduction in ocean biomass content of up to 80% [54]. The proportion of fish stocks within biologically sustainable levels decreased from 90% in 1974 to 62.3% in 2021 [22]. Furthermore, recent assessments have presented worrisome statistics regarding fish and shellfish stocks. Out of the 397 commercially exploited stocks, a staggering 69% were classified as overfished, reflecting the consequences of excessive fishing activities. Additionally, 51% of the stocks were found to be operating beyond safe biological limits, putting their long-term sustainability at risk. These findings become even more concerning when considering that only 12% of the stocks met the established guidelines outlined by the Common Fisheries Policy (CFP), which governs fisheries in the European Union seas [26]. Such data emphasizes the need for immediate measures to address overfishing and ensure the implementation of sustainable practices for the preservation and future of these vital marine resources.

Portugal has Europe's 5th largest exclusive economic zone (EEZ), the 3rd largest of the EU, and the 20th largest EEZ globally, at 1,727,408 km2. The Autonomous Region of Azores (Portugal), a group of 9 islands spread over 600 km wide, is a major contributor to the size of the country's EEZ. In the Azores, fisheries are considered artisanal and small-scale in nature, with a multisegmented fleet, targeting multiple species with a wide range of fishing gears or metiers [12], managed under the CFP and by national and regional management policies [3]. Its annual landings in the last ten years are about 11.000 tons and valued at about 33M€.

The Region implemented many technical and spatial measures, monitoring scientific surveys and the fishing effort, biological data, or the size composition of the species landings, are regularly collected under the national DCF (the PNRD). Despite efforts, information is usually deficient or inadequate for most species. Among the 138 species that landed in the region between 2009-2019, twenty-two (18 fishes, two mollusks, and two crustaceans) were selected as priority stocks according to the FAO and ICES criteria [56]. Most of these showed a decreasing trend in their abundances, and only four stocks are currently assessed using data-limited approaches [56]. However, the uncertainty inherent to data-limited stock assessments, such as the existing ones, can compromise the ability to inform management [18, 25, 33]. In addition, the more robust population assessment methods require considerable data, which are virtually impossible to obtain for all species, and limited fisheries data [34], such as our study case.

Monitoring the size of species over time is crucial for understanding fishing impacts. Fishing reduces population sizes by removing older and larger fish, disrupting density-dependent regulations, and leading fish to adjust life history traits like reproduction and growth [53]. This can indirectly affect predators, prey, and competitors, often resulting in a more even distribution of species, though heavy fishing can reduce species numbers. However, manually measuring large daily datasets across all species is impractical.

Therefore, tracking dynamic changes in fisheries practices over the years is an important and complex task. To enhance these studies, the main objective of this dissertation is:

Provide and study tools to improve data collection and analysis that assist in decision-making regarding the preservation of marine resources.

In order to achieve our goal we divided the work in 3 steps resumed in Figure 1.1:

1. The first was an analysis of the daily landing surveys conducted by OKEANOS, through a Network Analysis (published: EPIA Conference on Artificial Intelligence).
2. Then, we explored methods for reconstructing missing data regarding the fishing gear used in these surveys (under revision: Discover Data, Springer).
3. Finally, we used the FishMetrics system, which provided us with access to fish images and measurements to study different methodologies for automating the measurement of various species in real time (in preparation for submission).

The dissertation will begin by presenting the background information necessary to understand the key concepts used in this work in Chapter 2. Then, in Chapter 3, we will introduce the two datasets. Using the first dataset, we will conduct an in-depth network analysis. Next, we will describe the methodologies applied for both data reconstruction and size prediction. Finally, in Chapter 5, we will analyze the results of the data reconstruction and size prediction methods.

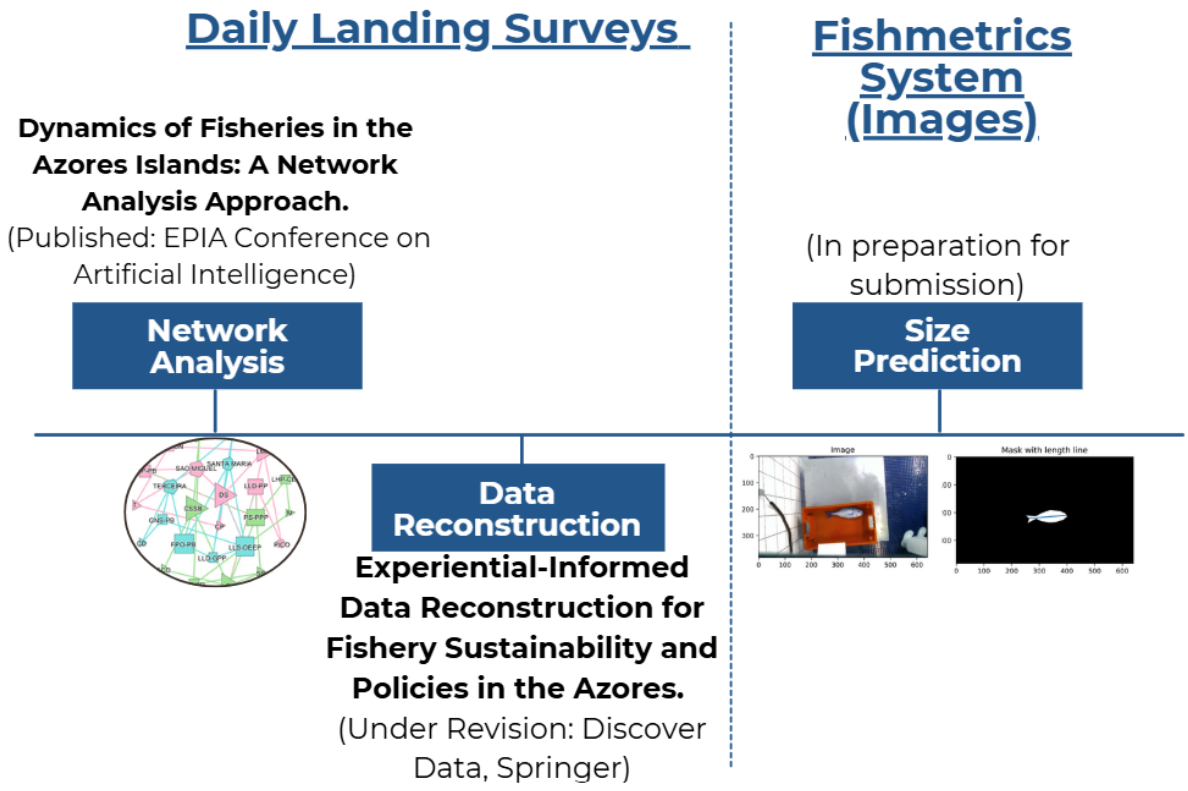


Figure 1.1: Summarization of the steps of the work.

Chapter 2

Related Work

Aquaculture science has made significant progress in response to the growing demand for food, but several challenges remain. A key issue is incomplete datasets, which can lead to distorted conclusions and incorrect decisions in research. Additionally, the high volume of daily fish landings makes manual annotation of fish nearly impossible, resulting in overlooked information. Current approaches used by researchers to address these problems in similar contexts use machine learning algorithms for data reconstruction, while deep learning has proven effective and accurate for real-time fish size estimation from images. As automation and artificial intelligence continue to advance, machine learning is becoming essential in aquaculture, creating new opportunities for digital fishery management [69].

2.1 Data Reconstruction

Data reconstruction has become crucial in many fields due to its importance in recovering missing information. Machine learning is a powerful tool for spatial reconstruction, with the potential to improve accuracy and efficiency in various applications. As machine learning algorithms continue to evolve, spatial reconstruction is expected to become more precise and widely adopted, with supervised models being the most commonly employed for these tasks.

For example, in a study on gene network reconstruction [61], microarray experiments were used to identify gene interrelations. A supervised learning approach, utilizing Decision-Tree-related classifiers [62], was employed to predict gene expression based on other gene expression data. Similarly, Liu et al. [43] tested machine learning models such as SVM [63], Random Forests [10], Gradient Boosting [15], and Deep Neural Networks (DNN) [69] for geomagnetic data reconstruction. In another study, Nurdjaman et al. [49] used Neural Network [69] to reconstruct sea surface temperature data.

In the work of Lafolla et al. [35], the authors take a step further by incorporating feature engineering alongside machine learning models to reconstruct sea wave data. Feature engineering, which involves extracting and transforming variables from raw data, plays a crucial role in

defining the input to a machine learning algorithm and significantly impacts its performance. In their study, they added power spectral density (PSD) information, which was not present in the raw data, to improve the machine learning model's input. Additionally, they discovered that records of the PSD could carry valuable information for evaluating the current sea state. As a result, they fed the machine learning model with the last five PSD records, enhancing its predictive capabilities. The machine learning models tested were Decision Tree, Bagging, Boosting, Random Forest, and Convolutional Neural Networks (CNN), explained later in this section.

However, none of these works address the issue of data imbalance, which remains underexplored in data reconstruction problems. Currently, classification solutions for imbalanced datasets are primarily focused on algorithm-level and data-level approaches. At the data level, both oversampling and undersampling strategies are employed to achieve data balance through data reconstruction [66].

Random undersampling (RUS) methods remove examples from the majority class to balance the dataset. The drawback of RUS is that it can lead to the loss of potentially useful information, which may be crucial for classification, thereby reducing the model's performance [37]. On the other hand, random oversampling (ROS) techniques, also known as upsampling, address dataset imbalance by replicating examples of the minority class. While ROS does not result in information loss, it can lead to overfitting and introduce additional computational costs, especially if the imbalance ratio is high [37].

Table 2.1 resumes the work references for data reconstruction used in this thesis.

Table 2.1: References resume for data reconstruction

Reference	**Dataset**	**Models**	**Feature Engineering**
Soinov et all.	Microarray datasets of genes	Decision Tree related classifiers	
Liu et al.	Geomagnetic data	SVM, Random Forest, Gradient Boosting and DNN	
Nurdjaman et al.	Surface temperature data	Neural Networks	
Lafolla et al.	Sea wave data	Decision Tree, Bagging, Boosting, Random Forest and CNN	Added power spectral density (PSD) information

2.2 Fish size prediction

Deep learning models have been extensively applied to various aquaculture tasks, such as identifying fish in underwater environments for counting or classification [1, 4, 42], and size estimation [2, 17, 28, 47, 67]. Fish size estimation is typically approached as a segmentation problem followed by a regression problem.

2.2.1 Deep Learning

Deep learning is a crucial area of research in Image and Video Processing, Computer Vision [39, 58], and Bioinformatics, among others. This has led to the development and application of various deep-learning techniques within these fields. Generally, deep learning techniques are categorized into three main types: CNN, Restricted Boltzmann Machines (RBM), and Autoencoders. In this section, we focus on CNNs, as they form the foundation for the models used in this thesis.

When it comes to substantial training of multiple layers, the CNN is considered the most momentous approach for a variety of applications [41]. The CNN architecture shown in Figure 2.1, consists of mainly three types of layers, namely: convolutional layers, pooling layers, and fully connected layers. Training the network can be divided into forward and backward stages. In the forward stage first, we classify the input image depending upon its weights and bias for each layer. The loss cost is calculated from the input data by using the predicted output. In the backward stage, depending on the loss cost measured, the gradients are calculated for each parameter. Using the gradients, it then updates the parameters for the next iteration. The training procedure can be halted after a satisfactory number of repetitions. The functionalities of the said network are described as follows:

1. **Convolutional Layers**: In this layer, the CNN uses numerous filters to convolve the entire image including intermediate feature maps and generating different feature maps [60].

2. **Pooling Layers**: This layer is similar to the convolution layer but minimizes the measurements of feature maps and also the parameters of the network. Generally average and max pooling are used. For average and max pooling, Boureu et al. [8] have demonstrated the theoretical details of their performances. Among all the three layers, the pooling layer happens to be the most profusely investigated. There exist mainly three approaches, with varied usage, that are related to the pooling layers.

3. **Fully-connected layers**: The final layer of CNN, a Fully Connected (FC) layer, also known as a dense layer, is a type of layer used in Neural Networks where each neuron or node from the previous layer is connected to each neuron of the current layer.

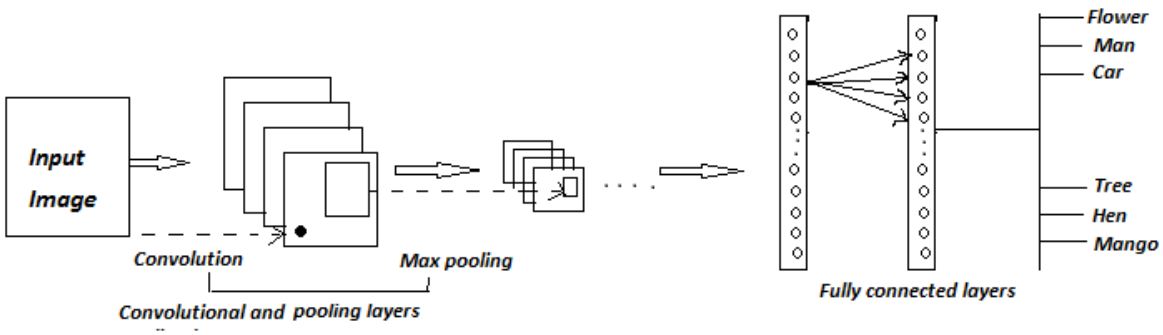


Figure 2.1: Pipelined Architecture of Convolutional Neural Network (**source:** Sinha et al. [60])

2.2.2 Fish Segmentation

Deep learning has seen significant application in computer vision tasks such as image classification, object detection, image retrieval, and image segmentation.

Object detection or localization extends image classification by not only identifying the classes of objects in an image but also determining their locations, typically using bounding boxes or centroids. Semantic segmentation aims for a finer level of inference by labeling each pixel in an image according to the object or region it represents. Instance segmentation goes a step further by assigning different labels to individual instances of objects within the same class, effectively combining object detection with semantic segmentation [14, 30].

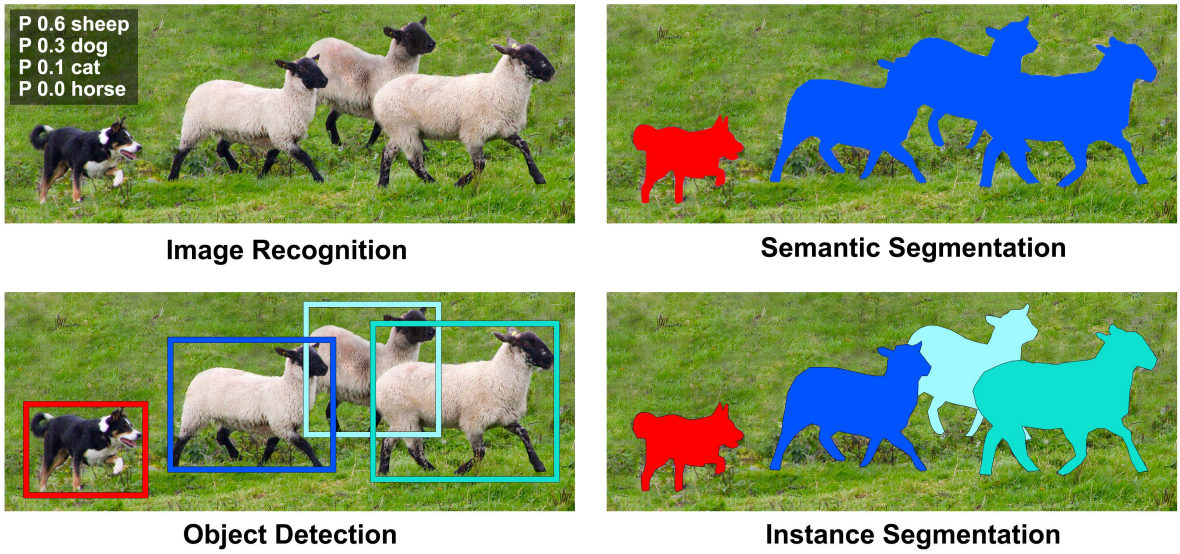


Figure 2.2: A sample annotated image illustrating the difference between image-level annotations, object-level annotations, and segmentations at the class/semantic- or instance-level. (**source:** Tedrake [64])

The segmentation may involve segmenting the entire image, or just specific parts like the head and tail [2, 50], or the entire fish [17, 28, 47]. Given that images often contain multiple fish, the problem can be treated as an instance segmentation problem. Several state-of-the-art instance segmentation algorithms have been developed, specifically Mask R-CNN (Mask Region-bases Convolution Neural Model) [32], YOLACT++ (You Only Look At Coefficients) [70], SAM (Segment Anything Model) [36] and YOLO (You only look once) [52].

Mask R-CNN is commonly used for instance fish segmentation [2, 28, 47]. It extends the Faster R-CNN architecture by adding a branch for predicting segmentation masks, in addition to its object detection capabilities. This makes Mask R-CNN suitable for instance segmentation, where the goal is to detect objects and precisely segment the pixels corresponding to each object instance [6]. However, it is computationally expensive, and enhancing the model's robustness, such as through data augmentation, is necessary before training.

YOLACT++ is a real-time instance segmentation model that efficiently balances speed and accuracy by generating prototype masks and combining them with coefficients to predict object masks, making it ideal for applications requiring fast processing like autonomous driving and video analysis. Its simple architecture allows easy deployment and fine-tuning. However, it may have lower accuracy in complex scenes with overlapping objects compared to models like Mask R-CNN, and it requires substantial labeled data and computational resources for training.

SAM is a new AI model from Meta AI designed to "cut out" any object from any image with a single click. It's a promptable segmentation system with zero-shot generalization, capable of segmenting unfamiliar objects and images without additional training. However, it relies heavily on bounding box selection for segmentation, which adds one more step to the training process, turning it more complex.

YOLO has emerged as a promising and effective solution for object detection and segmentation [50, 68]. Known for its fast detection speed and high accuracy, YOLO outperforms traditional models like R-CNN [68]. It divides an image into grids, predicting objects' presence, class, and position in each grid. By using a CNN to extract image features and fully connected layers to predict object positions and classes, YOLO performs detection and localization within a single network, reducing redundant computation and speeding up the process [68]. The YOLO series has evolved to YOLOv9, which includes data augmentation techniques within its framework.

2.2.3 Length Estimation

There isn't a unique method for fish length prediction; different approaches are used depending on whether the environment is underwater or controlled. In underwater environments, obtaining ground truth values is extremely challenging. For instance, in the study by Muñoz et al. [47], they tested fish detection using YOLO, followed by segmentation with Mask R-CNN. For size estimation, they transformed pixel coordinates of key points, such as the snout and fork, into 3D measurements using stereoscopic vision system calibration and 3D triangulation. Fish lengths were then calculated as the 3D Euclidean distances between these points. However, this approach lacks ground truth length data in underwater settings.

Garcia et al. [28] processed stereo images acquired by the Deep Vision imaging system, directly placed in a trawl. After correcting for camera response nonlinearities, they used Mask R-CNN for fish localization and segmentation, refining the boundaries with local gradients. For length estimation, they applied morphological operations to extract the fish's skeleton, smoothed it with a closing operation, and refined it using the RANSAC [24] algorithm to fit a cubic polynomial. The fish's length was determined by finding the intersection points between this curve and the fish's outline, with the distances between corresponding points in the stereo images calculated using epipolar geometry. Like Muñoz et al., this method also lacks ground truth length data and is complex and costly.

Another approach that does not rely on segmentation is used by Yu et al. [67]. They established a Fish-Keypoints dataset and employed DNN for detecting fish and their key points. Utilizing a binocular camera system, they reconstructed a three-dimensional coordinate system to measure key points at the fish's head and tail, which facilitated the calculation of fish length. The study involved 1,576 fish images and corresponding annotations.

In contrast, Álvarez-Ellacuría et al. [2] utilized a more controlled environment by applying Mask R-CNN for unsupervised (fully automatic) European hake length estimation from images of fish boxes at the box level. Their approach involved first using the segmentation model to identify the fish head, then calculating the pixel length of the detected head and subsequently using a statistical model to convert this pixel length into centimeters. A second statistical model was then applied to estimate the total fish length. To measure performance, the computed lengths were compared with the annotated real lengths. The study was based on 42 images containing a total of 200 fish.

Also, Climent-Pérez et al. [17] analyzed 1,291 annotated images of fish trays from a small-scale fish market. They employed YOLACT for instance segmentation, using the resulting masks, bounding boxes, and species labels as input data for a machine learning model. To make the segmentation data suitable for regression, they converted the polygon representing the fish into a single numerical value by calculating the enclosed area. Additionally, to account for varying angles and distances in the images, they transformed the bounding box using the initial (x, y) point along with width and height, rather than relying on both initial and final coordinates, ensuring more consistent real-world size estimation. They tested seven machine learning models for regression, including XGBoost and SVM.

Table 2.2 resumes the work references for fish length prediction through images used in this thesis.

Table 2.2: References resume for fish length prediction using images

Reference	**Environment**	**Fish Detection**	**Length Estimation**	**Limitation**
Muñoz et al.	Underwater	YOLO and Mask R-CNN	3D Euclidean distances	No ground truth
Garcia et al.	Underwater	Mask R-CNN	Cubic polynomial	No ground truth
Yu et al.	Underwater	DNN	Measure key points at the fish's head and tail	Complex methods to find the points and calculations
Álvarez-Ellacuría et al.	Box	Mask R-CNN	Statistic	Relies on specific parameter distributions, difficult to generalize
Climent-Pérez et al.	Box	YOLACT	Polygon enclosed area and bounding box using its coordinates and dimension. Them applied ML models	Proposed system relies mainly on uncalibrated images

Chapter 3

Data Presentation and Analysis

In this section, we introduce the fundamental datasets and their analysis that support our study.

3.1 Datasets

Our primary data sources include (1) the LOTAÇOR/OKEANOS-UAc daily landings dataset, derived from the PNRD/OKEANOS-UAc inquiries database [44] and (2) the images and annotations collected in the Fishmetrics/OKEANOS database.

3.1.1 Daily landings dataset

The inquiries are systematically collected by samplers during fishery landings in the Azores' 9 main fishing harbors, that belong to 5 different islands. Figure 3.1 presents an illustration of the position of islands with the total weight of fishes caught associated with each of the islands (labels without background) and the number of landings registered for each harbor (labels with white background). The inquiries encompass a wealth of information, including the precise locations (island and harbor) of the landings, the common names and major species groups of the captured fish, the weight of the catch, the types of fishing gear employed, and other essential vessel-related details described in table 3.1. Each individual observation within these datasets is uniquely identified and timestamped. Spanning the years from 2010 to 2017, our data comprises a total of 30,281 observations and comprises 13 categories of fishing gears – **referred to as metiers**, detailed in Table 3.2. Table 3.3 provides an overview of the 13 classification types utilized to categorize the fish in our data, as well as the total number of landings, the total weight of fish caught and the average weight[1] for each classification, in kilograms (Kg).

However, it is crucial to recognize the inherent complexity of this data, encompassing both

[1]The calculation of average weights involved the most encountered fish species within each group, whose combined weight contributed to 90% of the total group weight. We obtained the necessary weight data using the *rfishbase* resource [7].

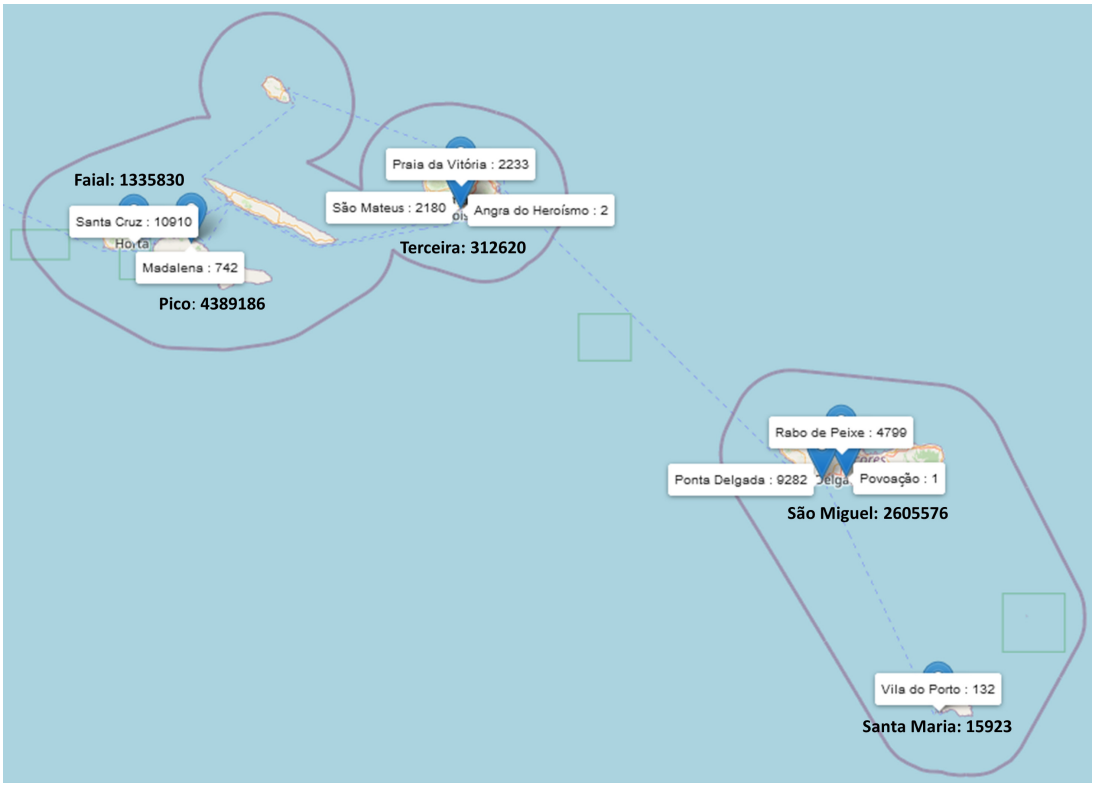


Figure 3.1: Azores harbors and the respective number of landings

Table 3.1: Features description

Feature	Description
Id_landing	Daily vessel landing Id
Data_landing	Date of the landing
Ilha	The island where the landing took place
Porto	Fish port of the landing
Vulgar	Common fish name
Classificação	Major fish species group
Peso	Weight of the landing per species
Matricula	Boat registration code
Classe_com	Boat size group code
Compff	Boat length
Metier	Type of fishing gear

univariate and high-dimensional time series analysis. To conduct a thorough analysis, we began with a focused study on métier and species classification, examining the proportion of major fish groups' weights caught for each métier over the years. This was followed by a broad visualization by creating cross-fisheries participation networks that link various fishing gears, fish species, and catching islands. Both analyses enabled us to explore how fishers have redistributed their efforts across different fisheries over time.

Table 3.2: Metier description, the total weight of fish caught and the average weight for each classification, in kilograms (Kg)

Metier	**Description**	**Landings Amount (Units)**	**Total Weight (Kg)**
LHP-TUN	Pole-and-line for tuna species	1002	6130088
LLS-PD	Bottom longline	10396	1203759
PS-PPP	Purse seine nets for small pelagic fish	2362	480957
LLD-PP	Deep-water drift bottom longline	115	293457
LHP-PB	Bottom fish handlines	11247	277624
LLS-DEEP	Deep-water non-drifting bottom longline	807	134733
LHP-CEF	Handline jigging for catching squids	3148	103761
GNS-PB	Coastal gillnets	943	17616
LLD-GPP	Surface drifting longline for large migratory pelagics	42	13595
PS-PB	Lifting nets for small coastal fishes	84	1375
LHP-PBC	Pole-and-line and coastal trolling for small pelagics	36	1071
FPO-PB	Fish traps	80	1014
FPO-CRU	Crustacean traps	2	53
NEI	Not Identified	17	34

Table 3.3: Major fish classifications, total number of landings, total fish weight and

Classification	**Landings Amount (Units)**	**Total Weight (Kg)**	**Average Weight (Kg)**
Tunas (T)	998	6131140	115.30
Continental Shelf Slope Demersals (CSSD)	6298	672199	57.92
Small Pelagics (SP)	2886	505364	2.09
Deep-Sea Species (DS)	2028	474636	3.56
Continental Shelf Slope Benthopelagic (CSSB)	5247	432526	4.00
Demersals (D)	3085	215418	79.10
Mollusks (M)	3181	104611	1.50
Coastal Demersals (CD)	5270	72392	7.55
Large Migratory Pelagics (LMP)	141	25019	650.0
Coastal Pelagics (CP)	877	23561	13.55
Small Coastal Demersals (SCD)	158	1605	0.30
Other Spp (OS)	110	651	NA
Crustaceans (C)	2	13	1.50

3.1.1.1 Data Analysis

To gain a comprehensive understanding of each species' dynamics and the fishery industry's overall sustainability, it is important to know the species' total catches (total weight) and how they change over time. This information is presented in Figure 3.2. It is evident in this case that the total weight of fish caught has decreased over the years, indicative of a potential intensive fishing or overfishing issue. The variations in catches are caused by various management practices, such as implementing Total Allowable Annual Catches (TACs), reducing fishing efforts, or changing the number of fishing vessels, among other factors. Therefore, it would be interesting to calculate catch-per-unit-effort (CPUE) using the effort hours. CPUE provides more reliable indicators of population abundance and is commonly employed in fish stock assessment models.

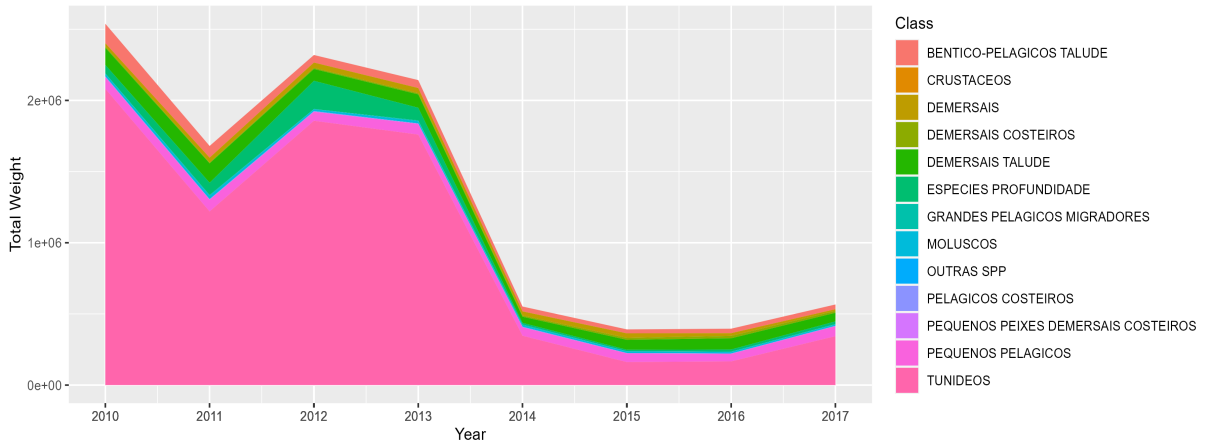


Figure 3.2: Total weight of each fish classification over the years

One important question in the fisheries industry study is the species caught across different metiers over time. Figure 3.3 shows the proportion of major fish groups weight caught for each metier over the years in the Azores, and a trend has emerged where significant species such as Tunas (*tunideos*) have declined in relevance, allowing other species to become more prominent. Tunas are migratory fish species, and catches may strongly fluctuate from one year to another, and the occurrence of different tuna species also varies mainly due to environmental causes.

The image in Figure 3.3 shows that the "Others" section represents the metiers that have a relatively small proportion in the overall fishery. To better understand the impact of these metiers on the less significant fish species, we have zoomed in on this section and presented the data in Figure 3.4. By analyzing the figure depicting the lesser-known metiers, we can observe a notable shift in representation. Some of these metiers become more prominent in the other image, disappearing from the current one. This observation suggests a potential improvement in the distribution of fishery activities among species in the latter image, indicating a more balanced and diversified utilization of fishing techniques

Mapping time series onto networks offers a solution by encapsulating intricate dependencies among constituent processes, including immediate and delayed interactions, as well as serial

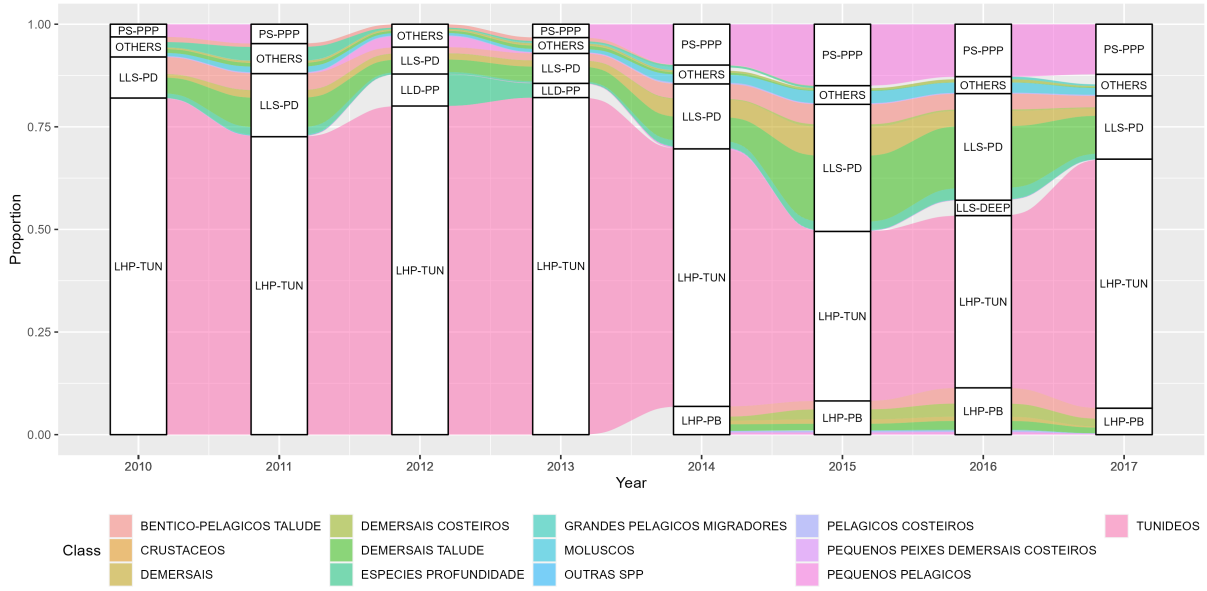


Figure 3.3: Distribution of major fish groups over the years per metier.

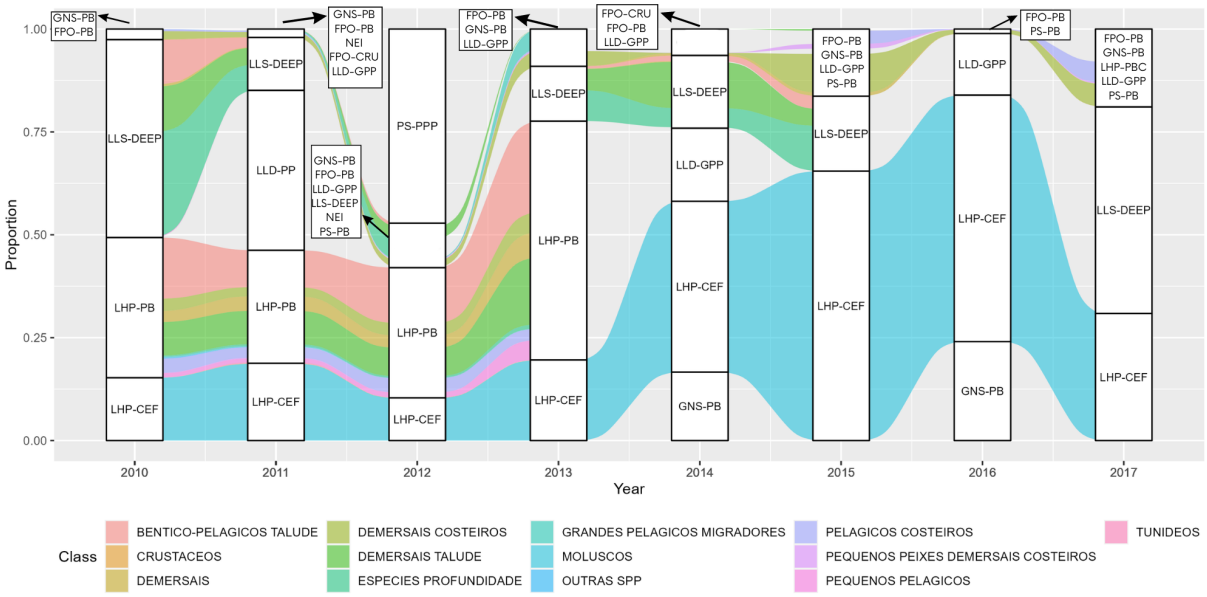


Figure 3.4: Distribution of major fish groups over the years per metier: small fish classes

dependencies [59]. Network analysis enables the assessment of important system characteristics by examining the network's topology [48]. Recent studies have applied network analysis to fisheries. For example, in [48], connectivity networks were constructed to describe how fisheries are linked through regional catch portfolios and timing of landings, interpreting the temporal structures of community fishing portfolios. In [27], networks for major ports were analyzed to identify systematic topological differences. Additionally, [21] used a network approach to understand by-catch structures in small-scale fisheries.

For this analysis, we performed several preprocessing steps to streamline our network analysis.

First, we identified a data gap from January 2014 to March 2014 and chose to fill it using data from the corresponding period in 2013, as it had a similar number of landings. However, due to significant differences in weight values between 2013 and 2014, we applied a scaling factor. This factor was calculated by comparing the total weights from April to December in both years, and we used it to adjust the 2013 data to estimate the missing weight values for 2014.

We also intentionally left out data from the harbors Angra do Heroísmo and Povoação, the classification Crustaceans, and the métier FPO-CRU due to limited data availability and low landing volumes. Additionally, the classification Other Spp and the métier NEI were omitted from our analysis because of their limited relevance.

After completing the imputation and data-cleaning steps, we generated the time series by aggregating observations. We calculated the mean value for each classification, métier, and island per month, then normalized each series, as shown in Fig 3.5. Harbors were excluded from this aggregation since they are closely tied to their respective islands and could introduce unnecessary complexity.

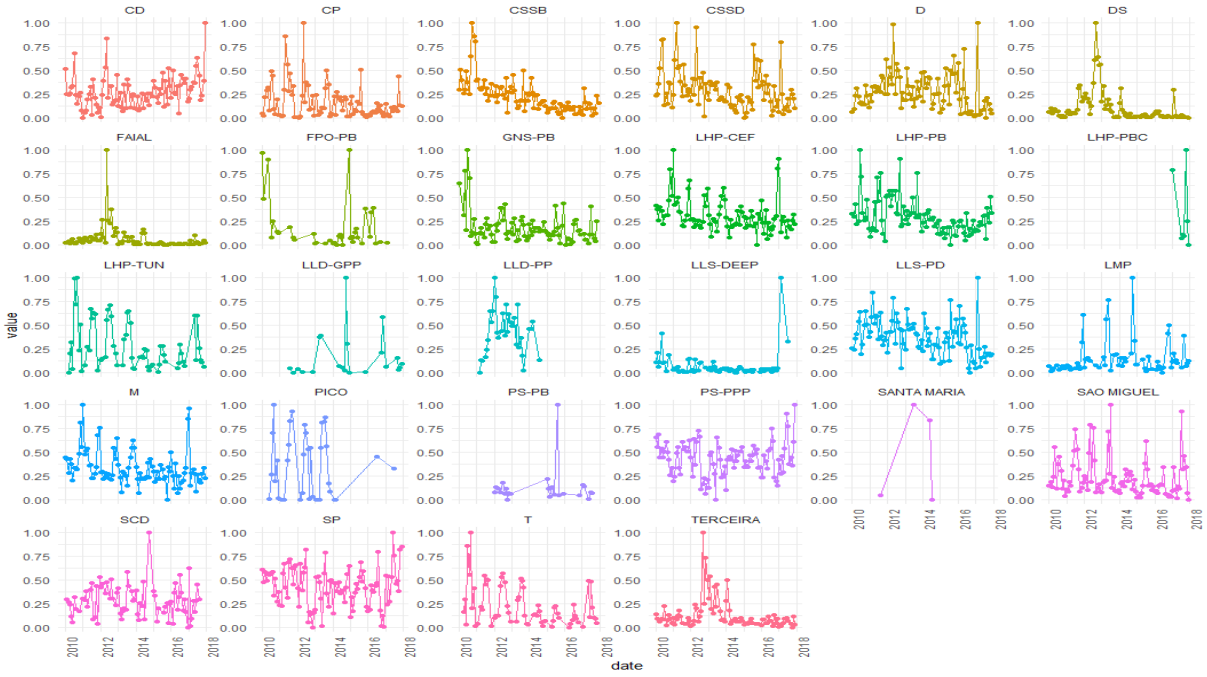


Figure 3.5: Time Series of Islands, Metiers and Classifications

Networks have become powerful tools for modeling complex systems and capturing interdependencies among time series data. Transforming time series into networks offers various approaches, including consolidating multiple time series into a single network or constructing a multi-layer network. In this context, our focus is on transforming a set of time series into a network.

A network or graph, represented as $G(V, E)$, is an ordered pair where V represents the set of nodes (or vertices) and E the set of edges (or links) between pairs of nodes belonging to V. A graph is most commonly represented as an adjacency matrix, denoted as A, in which an entry

$A_{i,j}$ equals 1, if there is an edge connecting the two nodes i and j, or 0 otherwise [59].

There are various approaches to constructing a network from a time series or a set of time series. We will be focusing on using a set of time series for this transformation, which works by mapping states of the time series into nodes of the network and creating links between those nodes based on a measure of distance or similarity [59]. This process begins with the computation of the distance between all pairs of time series, resulting in a distance matrix (D).

In this case, we will be using Dynamic Time Warping (DTW) [5] as the distance function, which aligns time series using a warping path, distinct from lock-step measures like Euclidean distance [5]. It optimizes the warping path to minimize the global warping cost, calculated through dynamic programming and a cumulative distance formula for two time series X and Y (both of length T), which is defined by Equation 3.1.

$$d_{dtw}(X,Y) = dtw(i=T, j=T) = \begin{cases} \infty & \text{if } i = 0 \oplus j = 0 \\ 0 & \text{if } i = j = 0 \\ \|X_i - Y_i\| + min \begin{cases} dtw(i-1, j) \\ dtw(i, j-1) \\ dtw(i-1, j-1) \end{cases} & \text{otherwise} \end{cases} \tag{3.1}$$

DTW stands out as a powerful technique for analyzing time series datasets due to its adaptability to variations and dynamic patterns, making it superior to rigid similarity measures. Its primary advantage lies in its invariance against shifting and scaling along the time axis. This unique feature has made DTW highly favored in pattern-matching tasks. Notably, DTW not only provides a distance measure between two sequences but also offers insights into how these sequences are aligned with each other. In certain cases, understanding the alignment can be as informative, if not more so, than the distance itself [11].

Network Construction: Following the computation of these distances, the next step involves converting the distance matrix D into an adjacency matrix A, that will represent the network. In our exploration of different network construction approaches, we tested:

- **k-Nearest Neighbors Network (k-NN):** each node is connected to the k other nodes with the shortest distances. This requires finding the k closest elements for each row i in D [23].
- **ϵ-Nearest Neighbors Network (ϵ-NN):** each node is connected to all the nodes whose distance is shorter than ϵ, a user-defined threshold [23].
- **Weighted Network:** is constructed by connecting all pairs of nodes and using their distances as weights. Typically, shorter distances correspond to stronger links, and the

weighted adjacency matrix can be defined as $A = 1 - D$ or normalized D_{norm} [23].

- **Networks with Significant Links:** connects nodes only if their distance is statistically significant [23]. For example, the significance of the Pearson correlation coefficient can be tested using the z-transformation.

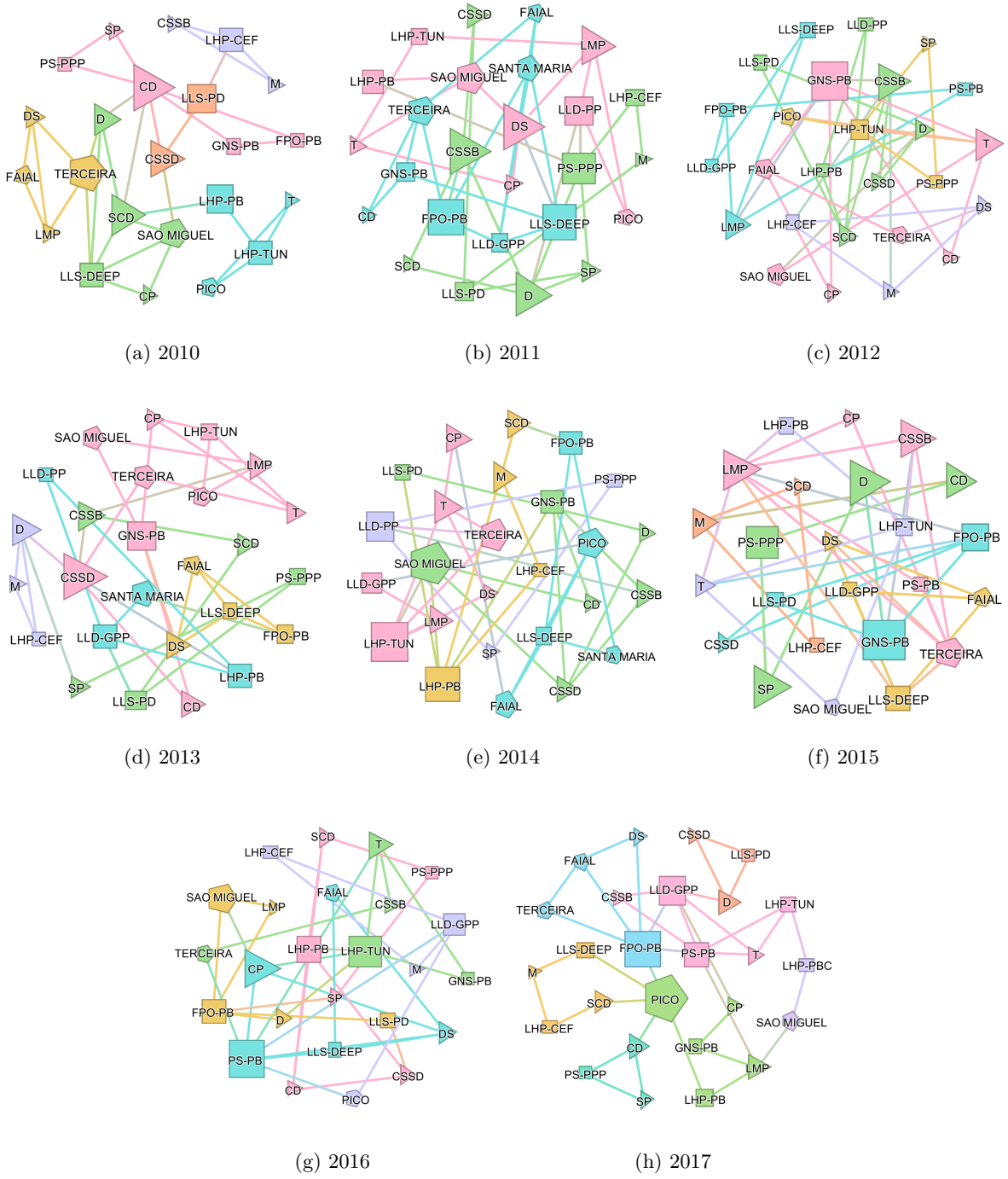
To transform our set of time series into networks, we employed the ***ts2net*** [23] library in R [29]. We considered various parameter values, specifically, different values for k (2, 3, 5, 7, and 10) and ϵ (0.3, 0.5, 0.7, and 0.9).

Although the network with significant links initially exhibited the highest mean modularity, a more dense network within communities, we ultimately selected the k-NN network with k=2 as our preferred choice. This decision was made because the network with significant links suffered from a lack of connections, resulting in a highly fragmented and disconnected network with a mean density of 0.02656478, while k=2 shows a mean density of 0.11386721, which is a sparse network, but not a fragmented one, being more suitable for the goals of our study.

Network Analysis: Networks can also be categorized as static or temporal. A static network uses the entire historical data to construct a single network, while temporal networks divide the data into time windows, creating multiple layers in a temporal (multilayered) network. Temporal networks enable the analysis of the system's evolution over time. For visualization purposes, we employed the igraph package [19]. In Fig. 3.6, the 2-NN networks for each year are depicted. Triangle-shaped nodes correspond to classifications, circle-shaped nodes represent islands, and square-shaped nodes denote metiers for better visualization. Nodes are color-coded to signify their communities each year, identified using the "cluster_walktrap" function from the igraph package. This function identifies communities based on random walks. The size of each node is according to its betweenness. The rate of new edges that emerge between years is significantly high, demonstrating a great change in the network over the years and translating into a significant shift in relations between metier, islands, and classifications.

Tunas emerge as a prominent contributor to the overall weight of fisheries in the Azores, as seen in Figure 3.2. However, as time has progressed, we've discerned a concerning decline in the total catch weight of the species. This decline and diminishing connections to once-consistent islands raises pressing questions. For instance, tunas exhibit strong links with São Miguel, Pico, Terceira, and other islands in various years. However, these connections seem to have diminished in recent years. As the connection with Tunas decreases, Large Migratory Pelagics appear to play a more pivotal role in São Miguel's fishing activities. It compels us to consider the sustainability of tuna populations and the possibility of shifts in their migration patterns. This underscores the paramount need for vigilant fisheries management in these regions.

Moreover, the vanishing connections involving various fish classifications and Faial in more recent years raise another layer of concern. Faial displayed connections to Deep-sea Species in recent years, contrasting with its past association with Coastal Pelagics. Meanwhile, Terceira



(a) 2010 (b) 2011 (c) 2012

(d) 2013 (e) 2014 (f) 2015

(g) 2016 (h) 2017

Figure 3.6: Networks over the years.

exhibits connections to Deep-sea Species and CSS Benthopelagics, although only during specific years, indicating variability in fishing patterns. They suggest a noteworthy alteration in the marine ecosystem around the island over the years.

Variations in fish catches are affected by management practices such as TACs, adjustments in fishing effort, changes in the number of fishing vessels, and reductions in fish sizes, among other factors. Given these complexities, CPUE calculated using effort hours offers a more reliable indicator of population abundance and is commonly used in stock assessments. Tracking fish sizes over time is also crucial for understanding population health, growth rates, and environmental impacts, which are key to sustainable fisheries management.

However, obtaining accurate CPUE data for commercial or recreational fisheries is challenging and often error-prone. Data is usually collected through survey sampling programs and extrapolated to estimate total catches, relying on statistical assumptions about the representativeness of samples related to species landings, fishing vessels, and fishing methods. Accurately linking fishing gear to specific fish landings to estimate effort is often limited, and fishing surveys frequently suffer from significant information gaps, particularly concerning fishing methods and efforts.

Tracking fish sizes poses additional challenges due to the time-consuming and labor-intensive nature of manual annotation, making it difficult to maintain consistent and comprehensive data over time.

Our work addresses these issues by developing methods to reconstruct missing data from existing Azores fisheries surveys and exploring techniques for automating fish measurements. These efforts aim to improve data accuracy for future stock assessments and enhance fish sustainability.

3.1.2 Fish Images

The fishmetrics system [45] automates fish data collection in fishing vessels and fish auctions and simplifies the fish measurement procedure by uploading the data to a remote database, where measurements can be taken at a later stage, not requiring physical access to fish. The system uses 3D motion information obtained with a 3D scanner to detect the presence of new fish boxes in conveyor belts or planar scales. Once a new box is identified, RGB and depth images are acquired, processed, and uploaded to the database. The database itself has a total of 78621 images encompassing information on more than 50 species from 2022 to 2024. An example of the measurements system is presented in Figure 3.7.

We collected a total of 1,000 measurements and associated data from the Fishmetrics/OKEANOS database, covering 453 images of four different fish species, spanning from 2022 to 2024. Each image contains several pieces of information, but the key variables relevant to our study are Measure ID, Length of fish (mm), box weight (kg), picture ID, when the picture was taken, species name, size of boat, boat, seaport, fishing gear used and post-distance annotation coordinates ('X1', 'Y1', 'Z1', 'X2', 'Y2', 'Z2').



Figure 3.7: Example of a sample in the FishMetric's system.

Table 3.4: Species annotations

Specie Name	Total Pictures	Total Measurements
Alfomsim	120	345
Goraz	148	211
Imperador	129	240
Peixao	56	204

The four species of focus are Goraz, Imperador, Alfonsim, and Peixão, selected due to their significant importance in fisheries. Table 3.4 provides a detailed count of images and measurements for each species, while Figure 3.8 shows the distribution of their lengths.

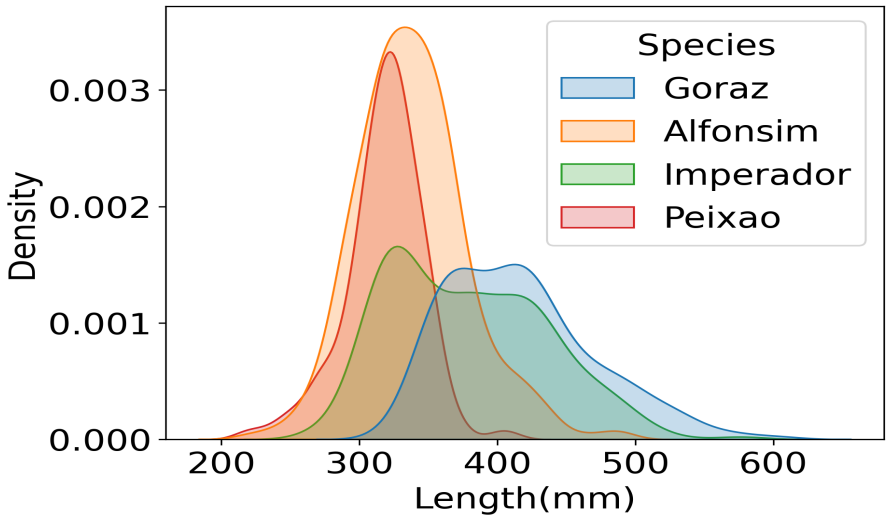


Figure 3.8: Distribution of length for each specie.

Chapter 4

Methodology

This section presents the methodology used in this thesis.

4.1 Data Reconstruction

Accurately predicting missing information in data analysis is a challenging and complex task with various applications in various fields. We explored five questions regarding the optimization and improvement of machine learning models on a dataset:

1. How does the performance of different machine learning methods vary on our dataset, and how can we identify the ideal time frame for making accurate predictions on our time series data? What impact does this have on the precision and stability of our models?
2. To what extent can incorporating contextual variables improve the accuracy and robustness of our predictions?
3. Can selectively remove features from our dataset enhance the performance of our models?
4. How does rebalancing an imbalanced dataset affect the quality and representativeness of our training data, and what are the trade-offs involved in different rebalancing approaches?
5. How can ensemble learning techniques be used to enhance the accuracy and robustness of our models, and what are the best practices for incorporating these methods into our workflow?

First, this study explores the effectiveness of several supervised machine learning methods in predicting the metiers of fish landings. Different training and testing evaluation periods are assessed to determine the optimal timeframe for prediction, aiming to find the right balance between the size of the training set and the prediction horizon to ensure accuracy and timeliness.

Additionally, the impact of incorporating context-related variables on the model's performance is examined, with the hypothesis that these variables can improve accuracy by capturing more complex relationships.

To enhance the model's performance further, the effect of removing redundant or noisy features from the dataset is investigated. The goal is to identify low-contribution features and remove them, while also reconstructing the dataset in a balanced manner to better represent both majority and minority classes.

Finally, the potential for improved performance through combining the best models from previous solutions is explored. Ensemble learning techniques are considered, where different models are combined to achieve more accurate predictions.

4.1.1 Methods

This section describes the methods used to address the research questions, including pre-processing, predictive solutions, and evaluation methods.

4.1.1.1 Pre-processing Solutions

Pre-processing strategies are crucial for enhancing model performance by refining the data used for training. Different approaches were adopted to improve model predictiveness through data preparation:

- **Context-related variables:** To investigate whether adding context-related variables can improve the performance of machine learning models, the study added information on the weight of each fish group classification over the past six months. This information was relevant since the weight of fish is likely to vary over time, and including it could help improve the accuracy of the models. To calculate the new variables, we first identified the date of each observation and then computed the maximum, minimum, and mean weight of each class over the past six months. We then added these new variables to our dataset, which increased the number of features available for modeling.
- **Feature Selection:** Feature selection is a technique used to improve the performance of machine learning models by removing noisy or irrelevant features from the dataset. In this study, the Boruta [40] algorithm was employed, an all-relevant feature selection wrapper algorithm that ranks the features according to their importance and determines which features are relevant, irrelevant, or uncertain.
- **Balanced Reconstructions:** Resampling strategies modify the original data distribution to meet specific user-defined criteria. The study evaluated the performance of different sampling techniques, including undersampling of the majority classes and oversampling of the minority classes from the UBL package [9], and two hybrid techniques that combined

both approaches, Importance Sampling [46] and SMOTE [13]. The functions and the corresponding parameters are described in Appendix A.

4.1.1.2 Predictive Solutions

Supervised learning models are commonly used in classification solutions [57]. Tree-based approaches are a popular class of algorithms [31]. Therefore, for our study, we have chosen three tree-based approaches: Decision Tree, Random Forest, and Boosting.

To ensure that our work is easily replicable, we used the implementations of these tools available in the free and open-source R environment. Concerning the parameter settings for each of these methods. We carried out a preliminary test to search for optimal parameterization (i.e., the setting that obtains the best possible results within a certain set of values of the parameters). The search for optimal parameters was carried out for each combination machine learning model - time window, described in the 4.1.1.3, and the results are detailed in Appendix B.

4.1.1.3 Evaluation Methods

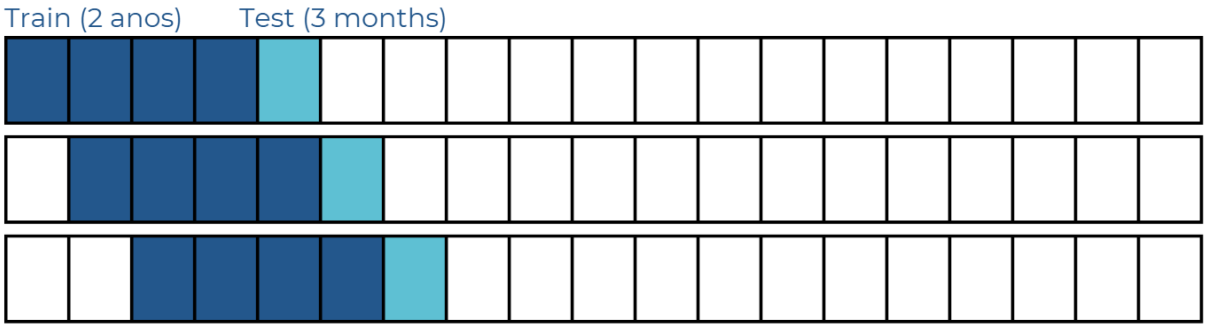
To evaluate the performance of predictive models on this dataset, it was decided to consider different time windows for training and testing. So it was decided to use different time windows for training and testing. We selected three methods, exemplified in Figure 4.1: sliding, growing, and full.

- **Sliding Window:** Trains the model on two years of data and tests it on three months, with the two-year period being before the three-month test period. We increment the window by one month for training and testing over the years;
- **Growing Window:** Trains the model on all the available data before the test period (which is also three months), starting with a two-year window and incrementing it by one month for both training and testing.
- **Full Window:** Both methods are commonly used in time series analysis, but as we needed to fill in missing values, we thought of using another approach called the full window method. This method considers all the data before and after the three-month test period, with increments of three months over the years for evaluation.

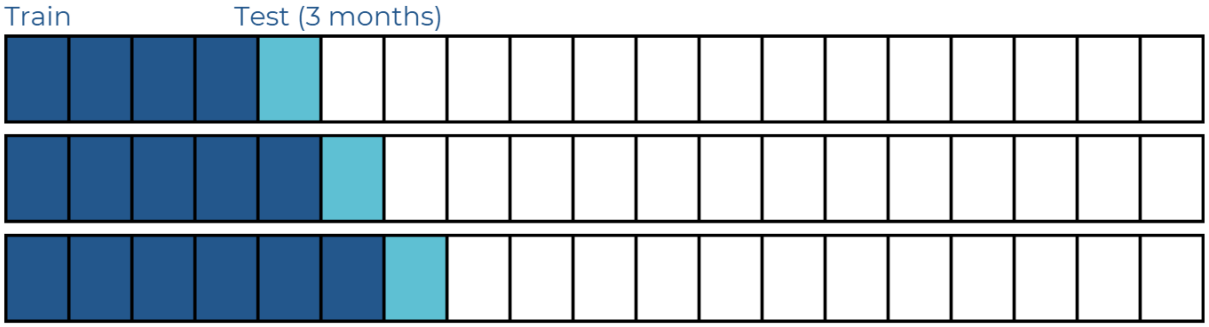
The most commonly used metric for classification evaluation is accuracy, which we employed for assessing performance across different time windows.

$$Accuracy = \frac{CorrectPredictions}{AllPredictions}$$

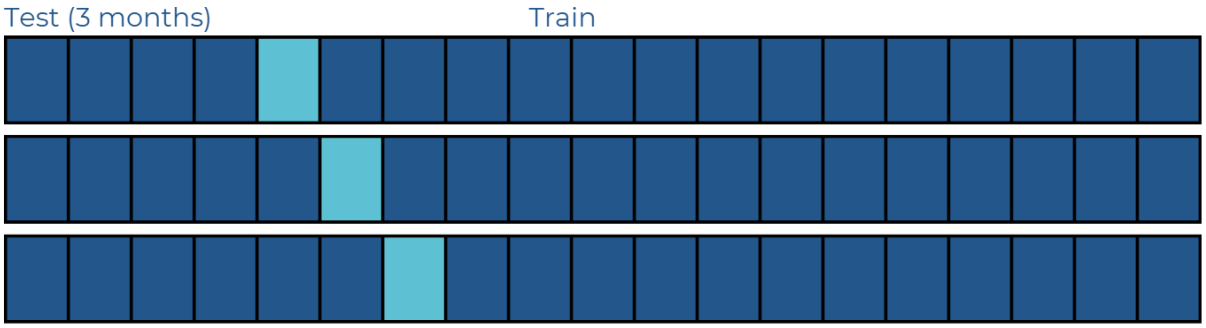
For the final evaluation, we employed accuracy along with the following metrics:



(a) Sliding Window.



(b) Growing Window



(b) Full Window.

Figure 4.1

- Geometric Mean: for multiclass classification can be used to evaluate model performance in a more balanced way across different classes, especially when dealing with imbalanced datasets. For multiclass classification, the geometric mean (G-mean) combines the true positive rates (or recall) of all classes to provide a single metric.

$$Recall_i = \frac{TruePositives_i}{TruePositives_i + FalseNegatives_i}$$

for each class i.

$$GeometricMean = \sqrt[N]{\prod_{i=1}^{N} Recall_i}$$

- Balanced Mean: is the average of each class weighted by its proportion in the dataset.

$$BalancedMean = \frac{\sum_{i=1}^{N} p_i * Acc_i}{\sum_{i=1}^{N} p_i}$$

- Imbalanced Mean: is the average of each class weighted by the inversion of its proportion in the dataset.

$$ImbalancedMean = \frac{\sum_{i=1}^{N} (1 - p_i) * Acc_i}{\sum_{i=1}^{N} (1 - p_i)}$$

4.2 Automated Fish Size Measurement

Given the methods introduced in Chapter 2 and computational limitations, we choose to use YOLOv9 for instance segmentation of the fish images. The ground truth masks for each image were created using the LabelMe tool [55] and then converted into the YOLO dataset format.

In some images, multiple fish are present. The annotated masks indicate the corresponding picture but not the associated measurement. To assign each mask to its respective measurement, we used the pixel length distance method explained later in this chapter 4.2.1. We ordered the observations by both measurement and pixel length distance, ensuring that the mask with the greatest pixel distance was assigned to the largest measurement. This process was also applied to the mask predictions to compare the ground truth with the predicted measurements.

Due to computational limitations, retaining the original pixel dimensions of the images was not feasible, which affected the computational effort for YOLO and length size estimation. Consequently, the images were resized from their original dimensions of 1080 x 1920 to 384 x 640 (height x width), respectively.

4.2.1 Pixel Length Distance

The pixel length of the line from the mouth to the middle of the tail of the fish was added as a feature in the dataset. This additional information can enhance model accuracy by providing more contextual detail. To determine this, we first needed to find the angle of the major intensity of the mask. We used image moments to compute the covariance matrix of the image, allowing us to identify the eigenvector associated with the largest eigenvalue, which corresponds to the major axis of image intensity.

An image moment is a specific weighted average of the pixel intensities, as described in Equation 4.1:

$$m_{i,j} = \sum_{x=1}^{X} \sum_{y=1}^{Y} x^i y^i f(x, y) \tag{4.1}$$

The center X and Y of the Image:

$$X = \frac{m(1,0)}{m(0,0)}$$

$$Y = \frac{m(0,1)}{m(0,0)}$$

The covariance matrix of an image can be defined as:

$$\begin{bmatrix} \mu^{'}(2,0) & \mu^{'}(1,1) \\ \mu^{'}(1,1) & \mu^{'}(0,2) \end{bmatrix} \tag{4.2}$$

Where:

$$\mu^{'}(2,0) = \frac{m(2,0)}{m(0,0)} - X^2$$

$$\mu^{'}(0,2) = \frac{m(0,2)}{m(0,0)} - Y^2$$

$$\mu^{'}(1,1) = \frac{m(1,1)}{m(0,0)} - X \times Y$$

Here, $m(0,0)$ represents the area of the image, $\mu'(2,0)$ represents the variance along the X-axis, $\mu'(0,2)$ represents the variance along the Y-axis, and $\mu'(1,1)$ represents the covariance between X and Y.

An eigenvalue/eigenvector decomposition of the covariance matrix reveals the principal directions of variation in the images. By computing the eigenvectors, we can identify the primary axes along which the image data varies the most.

Once the principal eigenvector is determined, it is used along with the center of the image to draw a line that passes through the center of the fish. Then, the points were examined along this line where the pixel intensity is different from zero, representing the fish's outline, as represented in Figure 4.2. Finally, the maximum distance between these points was measured, which gives us the length of the line from the mouth to the middle of the tail in pixels.

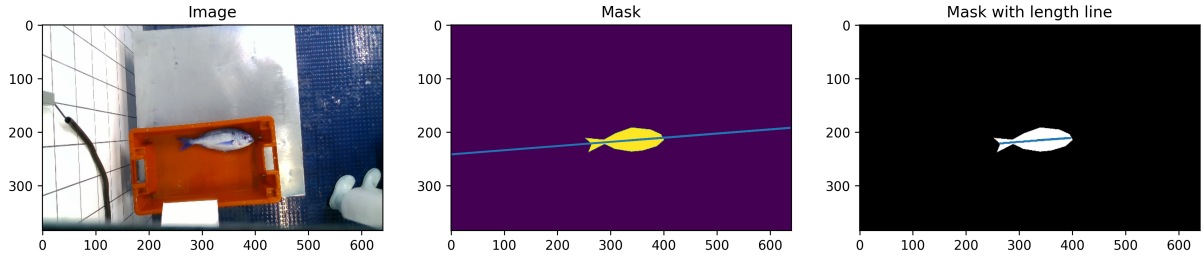


Figure 4.2: Visualization of Fish Length Measurement in Pixels.

4.2.2 Research Questions

To optimize the fish measurement, we explored 4 research questions for measurement prediction of four species:

1. How is the performance of YOLO as a segmentation model?
2. How does the performance of various machine learning methods differ for predicting the length of different fish species?
3. Which features are most strongly correlated with fish length, and does incorporating these features enhance prediction accuracy?
4. Is using only the pixel length sufficient for accurate length prediction, or is comprehensive mask information necessary?
5. Is a general approach more effective than species-specific training for fish length prediction?

The segmentation performance of YOLO was first assessed, exploring its effectiveness when trained with either a specific species or information from all species. Subsequently, various supervised machine learning methods were evaluated to identify the most effective approach for predicting fish length across different species. The analysis included examining image-derived features, such as the pixel length of lines, to determine their correlation with fish length and testing their impact on predictive accuracy.

Two prediction strategies were compared: one based solely on the mask's pixel length distance between key points on the fish and another utilizing full mask information via a CNN, as detailed in Section 2.2.1. The aim was to assess whether line pixel information alone was sufficient for accurate predictions and if the CNN model could better identify relevant characteristics.

Finally, the effectiveness of species-specific models was compared to a general model trained on data from all four species to determine which approach offered the most accurate predictions.

4.2.3 Methods

This section describes the methods used to address the research questions.

4.2.3.1 Machine Learning Models

To evaluate the predictions based on the pixel length of the fish, in addition to the CNN model, we tested four machine learning models: Linear Regression, SVM, XGBoost, and Neural Networks.

For implementation, we used the ***scikit-learn*** [38] library for SVM, Linear Regression, and Neural Networks, ***TensorFlow*** [20] for the CNN, and ***XGBoost*** [16] for the XGBoost model, using Python [51].

4.2.3.2 Feature Correlation

The correlation between the two observations was computed as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Here, x_i represents the i-th observation of variable x, and y_i represents the i-th observation of variable y. We also computed the p-value to assess the significance of each feature's relationship with the length, using the function *pointbiserialr* from the ***SciPy*** library [65] in Python [51].

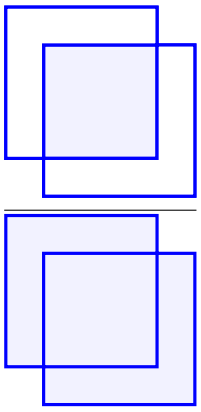
4.2.3.3 Evaluation

For each species, data was divided into 80% for training and 20% for testing, ensuring the data was shuffled. We performed 5-fold cross-validation using the training data to optimize regression models' hyperparameters. This was carried out separately for each species-specific dataset and the combined dataset containing all four species to create a general model. The hyperparameters tested for each algorithm are listed in Appendix C.

For segmentation, the metrics used to evaluate the performance of YOLO models are described as follows:

- **Intersection over Union (IoU):** measure that quantifies the overlap between a predicted bounding box and a ground truth bounding box. It plays a fundamental role in evaluating the accuracy of object localization.

$$IoU = \frac{\text{area of overlap}}{\text{area of union}} =$$



- **Precision and Recall:** Precision quantifies the proportion of true positives among all positive predictions, assessing the model's capability to avoid false positives. On the other hand, Recall calculates the proportion of true positives among all actual positives, measuring the model's ability to detect all instances of a class.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4.3}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4.4}$$

- **Average Precision (AP):** computes the area under the precision-recall curve, providing a single value that encapsulates the model's precision and recall performance.
- **Mean Average Precision (mAP):** extends the concept of AP by calculating the average AP values across multiple object classes. This is useful in multi-class object detection scenarios to provide a comprehensive evaluation of the model's performance.
- **mAP50:** Mean average precision calculated at an intersection over union (IoU) threshold of 0.50. It's a measure of the model's accuracy considering only the "easy" detections.
- **mAP50-95:** The average of the mean average precision was calculated at varying IoU thresholds, ranging from 0.50 to 0.95. It gives a comprehensive view of the model's performance across different levels of detection difficulty.

The evaluation metric used for regression was **Mean Absolute Error** (MAE), defined as Equation 4.5:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{4.5}$$

Chapter 5

Results and Analysis

The objective of this chapter is to analyze the results of the methodology described in Chapter 4, comparing the performance of different model development approaches for data reconstruction and fish length prediction. This analysis is crucial to understanding the effectiveness of each approach, while also identifying their respective strengths and weaknesses to inform the decision-making process.

5.1 Results Data Reconstruction

This section provides an overview of the answers to the research questions proposed in section 4.1. We aim to comprehensively understand the effectiveness of the different models and techniques used in this thesis.

5.1.0.1 ML models Performance Across Time Windows

The first research question addressed was the optimal combination of predictive solution and time window for training. Accuracy was analyzed for each predictive solution and time window method over the years, as shown in Figure 5.1. The results revealed that the full-time window produced more stable and improved outcomes across all models. Although both boosting and random forest models exhibited good performance, the boosting model with the full-time window was selected due to its consistent performance, stability over time, and capacity to handle complex variable relationships. This approach is referred to as the Base strategy.

5.1.0.2 Add New Features

To investigate whether adding new features could enhance the model's performance, a new dataset was created incorporating the features outlined in Section 4.1.1.1 and applied using the base strategy. Results revealed that the model with the additional features achieved the same

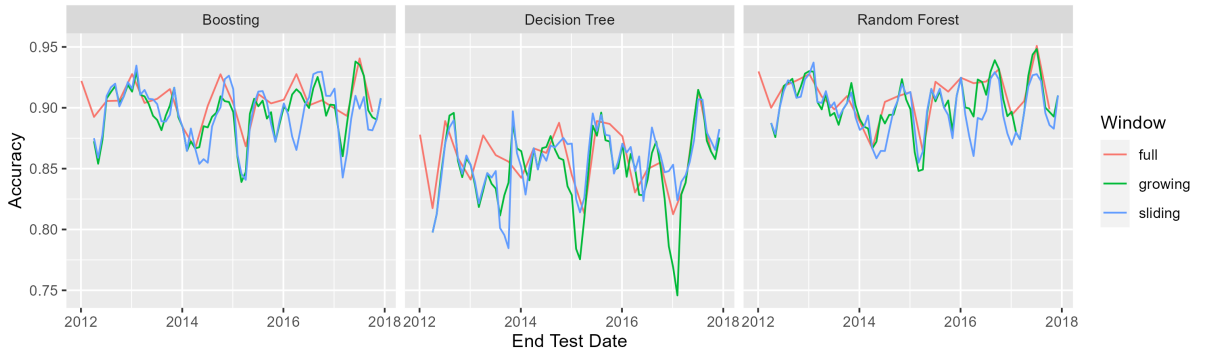


Figure 5.1: Machine learning methods and time windows

mean accuracy (90%) as the model without them. However, it is anticipated that including these features will significantly impact performance with a larger dataset. Therefore, the boosting algorithm, using the full window evaluation and the new features, was selected for further analysis. Results from these strategies applied to the new dataset are presented in Table 5.1.

Table 5.1: Accuracy of each metier class per method on the data set with new features

Metiers	**Methods**				
	Base	***ImpSamp***	***OverSamp***	***Ensemble***	***Feature Selection***
FPO-CRU	0%	0%	0%	0%	0%
FPO-PB	19,5%	50%	19,5%	**52,1%**	15,2%
GNS-PB	61,9%	56,7%	61,3%	**62,4%**	61,6%
LHP-CEF	**99,3%**	99,1%	**99,3%**	99,2%	**99,3%**
LHP-PB	**91,8%**	85%	91,3%	88,2%	**91,8%**
LHP-PBC	0%	**27%**	9%	17,9%	0%
LHP-TUN	**99,4%**	99%	**99,4%**	**99,4%**	**99,4%**
LLD-GPP	64,5%	48,3%	64,5%	**66.7%**	61,5%
LLD-PP	97,2%	92,6%	**98,1%**	94,5%	97,2%
LLS-DEEP	22,9%	78,7%	27,2%	**80.5%**	22%
LLS-PD	**86,6%**	72,4%	86%	71,8%	86,4%
PS-PB	0%	**62,7%**	6,9%	60,4%	0%
PS-PPP	96,6%	96,3%	96,6%	**96,7%**	96,6%
Final Accuracy	**91%**	87%	90,8%	88,2%	88,7%
Geometric Mean	**71.7%**	70.1%	49.1%	70.6%	69.8%
Balanced Mean	**91.02**	86.97	90.84	88.21	90.93
Imbalanced Mean	54.06	65.07	55.69	**66.80**	53.34

5.1.0.3 Feature selection

The new data set has 180 variables, but we suspected some might be affecting the model's performance. To investigate this, feature selection was performed as described in the 4.1.1.1 and ultimately identified 98 relevant variables, 76 irrelevant ones, and two uncertain variables. Only

98 variables were selected for our subsequent analysis.

In 5.1, it is noteworthy that the feature selection didn't improve any of the classes relative to the Base, and the final accuracy even decreased. This suggests that even the least relevant variables play an important role in distinguishing between classes.

5.1.0.4 Resampling

Table 5.1 reveals that certain minority classes, such as FPO-CRU, FPO-PB, LLS-DEEP, and PS-PB, presented in table 3.2, had low accuracy, with some even having an accuracy of zero. On the other hand, the majority of classes, including LLS-PD, LHP-CEF, and LHP-PB, demonstrated good performance, with an accuracy greater than 86%. In response to this imbalance, we constructed a more balanced dataset, focusing on these classes, except for FPO-CRU, which occurred only twice in the entire dataset.

We employed various strategies to select the best combination of parameters for undersampling and oversampling, as detailed in Appendix A. This resulted in two optimal models, OverSamp and ImpSamp, and their performances for each class are illustrated in Table 5.1.

The ImpSamp model was highly effective in improving the performance of the four proposed minority classes, as shown in Table 5.1. On the other hand, OverSamp showed only a slight improvement in LHP-TUN and LLD-PP, which were not our primary focus. These results suggest that balancing the data set positively impacts the performance of minority classes, albeit with a trade-off of slightly reduced accuracy in the majority classes.

5.1.0.5 Ensemble

After analyzing the results from the ImpSamp and Base models, the creation of a compensatory ensemble model was considered to maintain the accuracy of majority classes while preserving gains in minority classes. A new ensemble model was designed for this purpose.

Predictions were made using the Base and ImpSamp strategies within the Ensemble model. Results from the ImpSamp approach were retained if it predicted the FPO-PB, LLS-DEEP, or PS-PB metiers. If the ImpSamp approach did not predict any of these metiers, results from the Base approach were used.

The performance of the Ensemble model was evaluated and compared to previous models. As shown in Table 5.1, the Ensemble model preserved the accuracy achieved with the ImpSamp approach while maintaining performance for the majority of classes. This method achieved accuracy higher than 52% for all metiers except FPO-CRU and LHP-PBC, which only appeared in 2017 with very low occurrences.

5.1.0.6 Metric Evaluation

In addition to evaluating the accuracy of each classification strategy individually, we also compared their results using three different metrics: the geometric mean of the accuracy of each class, the average of each class weighted by its proportion in the dataset (balanced mean), and the average of each class weighted by the inversion of its proportion in the dataset (imbalanced mean). Table 5.1 showed that the Base strategy outperformed all metrics except for the imbalanced mean. This metric gives more weight to underrepresented classes, which had the best performance from the Ensemble strategy. As mentioned, the Ensemble strategy is particularly effective in predicting minority classes. Therefore, it can be concluded that the Ensemble strategy can provide the most accurate predictions for imbalanced datasets, while the Base strategy is more suitable for balanced datasets.

5.2 Discussion

It seems that the models' performance is very high for the most selective gears, those gears which are very specific to catch a small number of species, e.g., the LHP-CEF for catching squids, the LHP-TUN for catching tunas or the PS-PPP to catch small pelagics like the *Trachurus picturatus*.

Table 5.2 provides insights into the three iconic species associated with each class of fishery art, along with their respective percentages of total weight within that class. On the other hand, Table 5.3 displays the confusion matrix resulting from the ensemble method. By examining these tables, we can identify similarities between fishery landings and explore the misclassification patterns.

Table 5.2: Three Iconic Species of each metier

Metier	First	Second	Third
FPO-CRU	DEMERSAIS (93.94%)	CRUSTACEOS (6.06%)	
FPO-PB	DEMERSAIS COSTEIROS (52.64%)	DEMERSAIS (34.81%)	MOLUSCOS (5.76%)
GNS-PB	DEMERSAIS COSTEIROS (74.31%)	PELAGICOS COSTEIROS (20.31%)	DEMERSAIS TALUDE (2.09%)
LHP-CEF	MOLUSCOS (99.67%)	BENTICO-PELAGICOS TALUDE (0.11%)	DEMERSAIS TALUDE (0.09%)
LHP-PB	BENTICO-PELAGICOS TALUDE (36.92%)	DEMERSAIS TALUDE (23.03%)	DEMERSAIS COSTEIROS (16.48%)
LHP-PBC	PELAGICOS COSTEIROS (96.85%)	DEMERSAIS COSTEIROS (2.65%)	MOLUSCOS (0.50%)
LHP-TUN	TUNIDEOS (99.992%)	PELAGICOS COSTEIROS (0.003%)	GRANDES PELAGICOS MIGRADORES (0.003%)
LLD-GPP	GRANDES PELAGICOS MIGRADORES (96.52%)	DEMERSAIS (0.41%)	BENTICO-PELAGICOS TALUDE (0.38%)
LLD-PP	ESPECIES PROFUNDIDADE (99.879%)	DEMERSAIS TALUDE (0.119%)	OUTRAS SPP (0.002%)
LLS-DEEP	ESPECIES PROFUNDIDADE (51.16%)	DEMERSAIS TALUDE (30.76%)	BENTICO-PELAGICOS TALUDE (16.71%)
LLS-PD	DEMERSAIS TALUDE (46.79%)	BENTICO-PELAGICOS TALUDE (25.75%)	DEMERSAIS (15.52%)
PS-PB	PELAGICOS COSTEIROS (35.82%)	PEQUENOS PEIXES DEMERSAIS COSTEIROS (35.79%)	DEMERSAIS COSTEIROS (28.40%)
PS-PPP	PEQUENOS PELAGICOS (99.79%)	MOLUSCOS (0.11%)	TUNIDEOS (0.04%)

For example, an analysis of the data reveals that GNS-PB is predicted as LHP-PB in 35% of the cases. This can be attributed to the low occurrence of the GNS-PB fishery art in the dataset (as shown in Table 3.2), as well as the presence of common fishery species such as Demersais Costeiros and Demersais Talude. Similarly, LLS-PD is observed to be distributed between LHP-PB and LLS-DEEP. This can be explained by the fact that these three arts share

Table 5.3: Confusion Matrix generate with ensemble strategy

True	**Predicted**												
	FPO-CRU	FPO-PB	GNS-PB	LHP-CEF	LHP-PB	LHP-PBC	LHP-TUN	LLD-GPP	LLD-PP	LLS-DEEP	LLS-PD	PS-PB	PS-PPP
FPO-CRU	0	0	0	0	2	0	0	0	0	0	0	0	0
FPO-PB	0	24	0	1	16	0	0	0	0	0	3	2	0
GNS-PB	0	2	231	0	132	0	0	0	0	0	1	4	0
LHP-CEF	0	1	0	3377	20	0	0	0	0	0	3	0	0
LHP-PB	0	57	14	26	4100	18	3	1	0	74	288	64	4
LHP-PBC	0	0	0	2	21	5	0	0	0	0	0	0	0
LHP-TUN	0	0	0	1	5	0	1020	0	0	0	0	0	0
LLD-GPP	0	0	0	0	4	0	4	24	0	0	4	0	0
LLD-PP	0	0	0	0	0	0	0	0	103	6	0	0	0
LLS-DEEP	0	0	0	0	9	1	1	0	1	186	33	0	0
LLS-PD	0	8	0	3	331	0	5	1	2	438	2004	0	0
PS-PB	0	0	2	0	15	0	0	0	0	0	0	26	0
PS-PPP	0	2	0	22	15	0	1	0	0	0	0	12	1515

iconic fishery species like Demersais Talude and Bentico Peleagicos, leading to similarities in their predicted classifications.

Understanding the similarity between fishery arts is crucial for assessing the impact of fishing activities on fish populations and their habitats. Researchers can identify common trends and drivers of changes in fish stocks by analyzing catch composition, fishing techniques, and spatial distribution across different arts. This knowledge is essential for evaluating the sustainability of fishing practices and implementing measures to prevent overfishing and protect vulnerable species.

5.3 Results Length Prediction

This section provides an overview of the answers to the research questions proposed in section 4.2.2. We aim to comprehensively understand the effectiveness of the different models and techniques used in this thesis.

5.3.1 Segmentation

To determine whether a general segmentation approach, trained with information from all four species, is more effective than a species-specific model, we analyzed the results presented in Table 5.4. The table includes the metrics described in Section 4.2.3.3, which were applied to the test set. For mask predictions, a confidence level of 0.7 was used.

Table 5.4: Bounding Box and Segmentation Results using Yolo

				Box Metrics				Mask Metrics			
	Species	Images	Instances	Precision	Recall	mAP50	mAP50-95	Precision	Recall	mAP50	mAP50-95
General	Alfomsim	24	69	0.957	0.957	0.957	0.825	0.957	0.957	0.973	**0.773**
	Goraz	30	38	0.75	0.974	0.861	0.79	0.75	0.974	0.861	0.742
	Peixao	12	43	0.909	0.698	0.821	0.723	0.909	0.698	0.821	0.656
	Imperador	26	50	0.979	0.952	0.975	0.832	0.979	0.952	0.975	**0.804**
Specific	Alfomsim	24	69	0.918	0.971	0.96	0.777	0.932	0.986	**0.974**	0.738
	Goraz	30	38	0.99	1	0.995	0.857	0.99	1	**0.995**	**0.843**
	Peixao	12	43	0.875	1	0.968	0.832	0.875	1	**0.968**	**0.707**
	Imperador	26	50	0.98	0.997	0.991	0.822	0.98	0.997	**0.991**	0.774

For both bounding box and mask identification, the species-specific segmentation models generally demonstrate higher accuracy when considering mAP. However, for Alfomsim and Imperador, the mAP50-95 for bounding box identification is greater using the general model, indicating that the general model performs better in more challenging cases of bounding box detection. Additionally, for all species, the Specific model had the highest mAP, values for mask identification, while mAP50-95 achieves the highest for Alfomsim and Imperador with the general model and with specific for Goraz e Peixao.

To assign each predicted mask to the corresponding data observations, we sorted the

observations by both real measurement and pixel length distance, ensuring that the mask with the largest pixel distance was matched with the largest real measurement. We also addressed specific cases where the model failed to segment correctly: instances where the model detected zero fish in an image, cases where the model identified fewer fish than the actual number (e.g., detecting 2 fish when there were 3), and cases where it identified more fish than were present.

For the first and second scenarios, where there were no predicted masks associated with some observations, we assigned a pixel length value of zero and used an all-black mask for these entries in the dataset. In the final scenario, where the model predicted more fish than were observed, we sorted the masks by their confidence scores and discarded the ones with the lowest confidence, retaining only the number of masks that matched the actual count of fish in the image.

5.3.2 Feature Correlation

To understand whether adding more information to the regression models would enhance their performance, the correlation of each variable was explored. This involved transforming certain variables and incorporating additional ones related to the measurement data.

We split the information of when the picture was taken into "Month" and "Year" to better analyze if seasonal patterns or long-term trends have any correlation with fish length. The annotation coordinates ('X1', 'Y1', 'Z1', 'X2', 'Y2', 'Z2') were disregarded as they are derived from the post-distance annotation.

Also, the pixel distance length was included, along with the total number of annotations predicted associated with the same picture ID as the measurement. This was done under the assumption that knowing whether the fish is alone or in a group could enhance the model's performance. The correlations between these features and the total fish length are presented in Table 5.5.

Table 5.5: Features Correlation with Total Fish Length

Features	Correlation	pvalues
Length Pixels	0.767143	1.210992e-194
Total Annotations	-0.513224	2.752677e-68
Size of Boat	-0.498730	5.307472e-64
Year	-0.259774	6.928435e-17
Seaport	0.232366	9.997416e-14
BoxWeight	0.161277	2.940332e-07
Month	0.125652	6.768131e-05
Specie's Name	-0.077668	1.402149e-02
Fishing Gear	-0.034636	2.738507e-01
Boat	-0.034214	2.797326e-01

The pixel length shows a high correlation with the total fish length, around 0.76, with a

p-value close to zero, indicating it is a strong predictor. This is followed by the total number of annotations and the size of the boat, both having correlations around 0.5, with p-values also near zero. The variables Year, Seaport, Box Weight, and Month do not demonstrate a high correlation, but their p-values are still below 0.01. Meanwhile, Species Name, Fishing Gear, and Boat show low correlation with p-values greater than 0.01. The relevant features used in the next sections are in bold in Table 5.5.

For the test set evaluation, the Pixel Length and Total Annotations features were calculated using the results from the YOLO model.

5.3.3 Regression Models Hyperparameters Search

Our objective is to compare different modeling approaches: the general model trained on data from all species versus species-specific models, as well as machine learning regression models using pixel length information versus the CNN model utilizing mask information without pixel length. Additionally, we aim to assess whether adding the features described in the previous section enhances performance.

Following the hyperparameter search outlined in Section 4.2.3.3, Table 5.6 presents the best validation results for the general model with pixel length, species-specific models with pixel length, the general model using CNN, and species-specific models using CNN, along with the respective hyperparameters.

The table indicates that XGBoost generally performs better for regression models using pixel length as a feature, except for the species-specific model using Alfonsim information, where SVM achieved the best results. The top-performing models on the validation set incorporated the features described in the previous section, underscoring the importance of context in prediction. However, there were exceptions: the species-specific model with CNN using Peixão information and the general model with CNN did not benefit from these additional features.

5.3.4 Result on Test Set

We then evaluated the best models identified in the previous section on the test set. Table 5.7 shows the MAE for each model applied to this set. Notably, the general model outperformed the specific models in length prediction for Alfonsim, Goraz, and Imperador. However, for Peixão, the species-specific segmentation model achieved better results. These findings suggest that incorporating data from all species generally enhances length prediction accuracy.

Additionally, machine learning models using pixel length as a feature demonstrated superior performance, indicating the value of this feature for length prediction. In contrast, CNN models were less effective, as they did not capture the image's main characteristics as well compared to a simple method.

Table 5.6: Best results for hyperparameter search

		Specie	Model	With Features	Hyperparameters	MAE
Specific	Pixel Length	**Goraz**	XGB	True	Number Estimators: 100 Learning Rate: 0.1 Max Depth: 3	18.441041
		Alfonsim	SVM	True	C:1 Kernel: 'linear'	23.510021
		Imperador	XGB	True	Number of Estimators: 50 Learning Rate: 0.1 Max Depth: 3	21.737319
		Peixao	XGB	True	Number of Estimators: 250 Learning Rate: 0.01 Max Depth: 5	16.894008
	Mask	**Goraz**	CNN	True	Dense Layers: (100, 100) Learning rate: 0.01	52.538401
		Alfonsim	CNN	True	Dense Layers: (64, 64) Learning Rate: 0.01	34.808777
		Imperador	CNN	True	Dense Layers: (100, 100), Learning Rate: 0.01	54.310501
		Peixao	CNN	False	Dense Layers: (64, 64) Learning Rate: 0.01	56.823297
General	Pixel Length	**All**	XGB	True	Number of Estimators:100 Learning Rate: 0.1 Max Epochs: 3	21.684823
	Mask	**All**	CNN	False	Dense Layers: (100, 64) Learning Rate: 0.01	41.621822

Figures 5.2 and 5.3 provide a closer examination of the results by comparing the predicted and true length values for the General and Specific models presented in Table 5.7. The plots with titles in bold represent the best MAE results from Table 5.7. We found no evidence suggesting that empty masks affected the results, indicating that other features may play a significant role in improving prediction accuracy.

Table 5.7: MAE Results of Models in Test Set

Segmentation	Regression	Alfonsim	Goraz	Imperador	Peixao
General	ML	**21.206691**	**19.048979**	**14.420886**	17.613127
	CNN	37.459776	41.496488	49.052927	33.232810
Specific	ML	22.331382	20.435674	25.161522	**15.869960**
	CNN	42.435219	21.656772	56.337597	45.335283

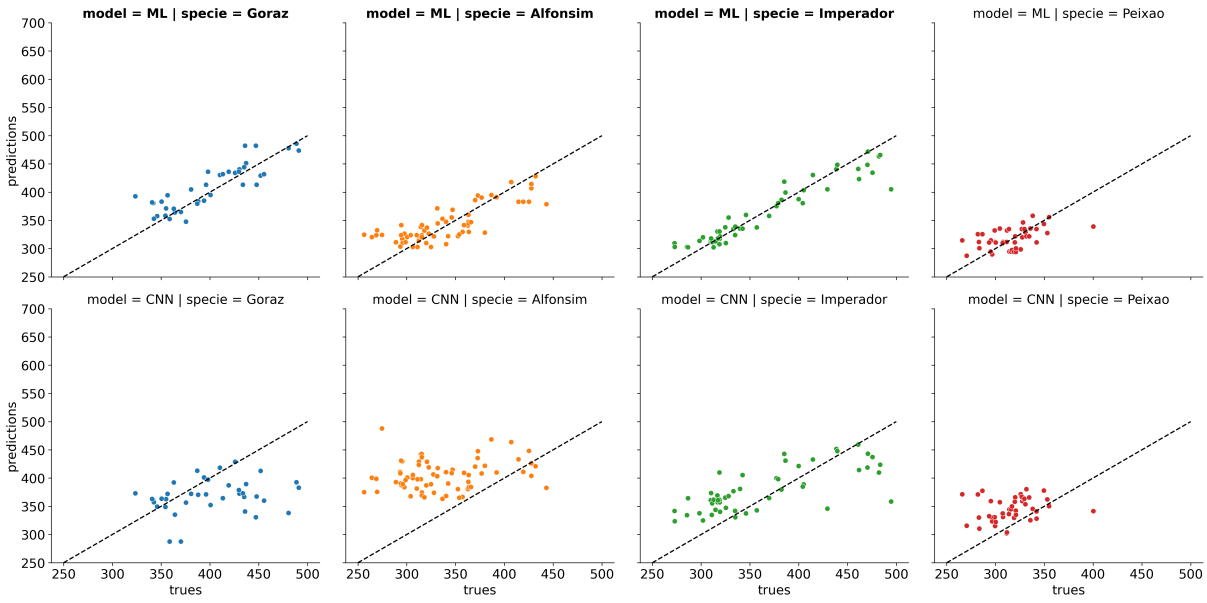


Figure 5.2: General Segmentation Model.

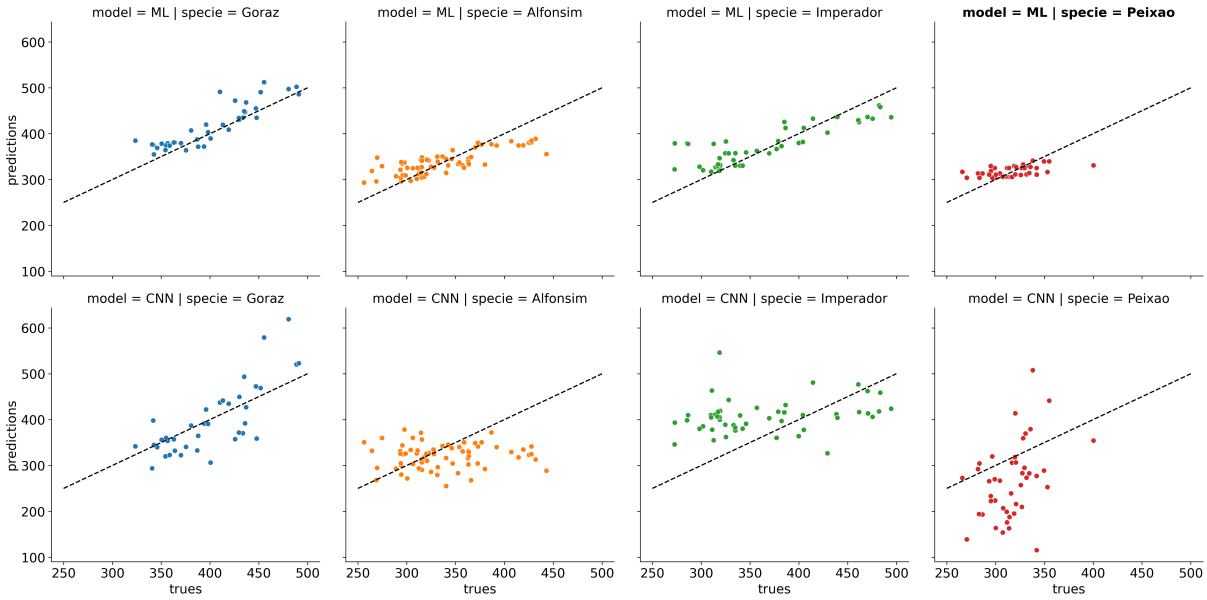


Figure 5.3: Specific Segmentation Model.

Chapter 6

Conclusion

In conclusion, this thesis addresses the critical challenges of data analysis and dealing with incomplete or missing information in marine science, which are essential for effective fisheries management and species sustainability. The tools explored in this work aim to provide reliable and comprehensive information and real-time analyses to support better decision-making in marine policies for ecosystem conservation.

Initially, we conducted an in-depth analysis of the LOTAÇOR/OKEANOS-UAc fishery landings dataset, using network analysis to better understand the dynamic changes over the years and their relationships with a wide range of factors, including islands, species classifications and metiers.

Next, we tackled the challenge of incomplete datasets in fisheries by applying machine learning techniques to fill gaps in the LOTAÇOR/OKEANOS-UAc dataset. Our work emphasizes the importance of predictive modeling in identifying misclassification patterns, highlighting critical species that require attention, and uncovering areas where inaccurate information may persist.

Subsequently, we explored methods for fish length prediction using computer vision and conventional machine learning models. Our findings demonstrated that extracting relevant information from models can significantly enhance performance, sometimes surpassing more complex methods. We also showed that segmentation is a crucial factor, with species-specific segmentation offering better results compared to models trained on diverse classes.

The significance of our research extends to actionable policy-making and real-time sustainability decisions. By providing decision-makers with a more comprehensive and reliable dataset, our work supports informed choices for resource management and conservation. Leveraging machine learning, we contribute to the sustainable management of fisheries and the preservation of marine ecosystems through evidence-based policies and real-time decision-making.

6.1 Limitations

While this work provides valuable insights, it is essential to acknowledge its limitations to understand its scope and identify areas for future improvement.

Firstly, the measurement annotations were not precisely matched to the mask annotations, which may have impacted the performance results. Additionally, the limited number of annotations per species restricted the training performance and model accuracy.

Furthermore, due to computational constraints, we were unable to explore various CNN parameters or test more computationally intensive methods, such as Mask R-CNN, which could potentially enhance performance. Similarly, image resizing was necessary due to computational limits, which may have reduced the pixel information available for accurate segmentation for pixel length estimation and CNN prediction.

6.2 Future Work

As part of future work, we aim to enhance the productiveness of our data reconstruction model. One promising approach is to incorporate Long Short-Term Memory (LSTM) networks. LSTMs are a type of recurrent neural network designed for sequence tasks, known for their ability to capture and utilize long-term dependencies in data. However, they require significant computational resources and have a complex training process.

Further exploration could involve including additional species and increasing the number of annotations to improve the accuracy of length predictions. Future research could also test more computationally intensive segmentation methods, such as Mask R-CNN and YOLACT, to evaluate their impact on prediction performance. Additionally, a more advanced approach could involve using YOLO to predict bounding boxes—given its demonstrated performance in this area—and then applying SAM for segmentation from the bounding box. This combination could lead to a more precise model.

Appendix A

Resampling Parameters in Data Reconstruction

In this thesis, we tested various functions for resampling techniques. We employed two methods to determine the optimal percentage of undersampling and oversampling for each class.

The first method involved setting the same percentage of undersampling and oversampling for all classes. The parameters tested are:

- **Undersampling Percentage**: {0.1 to 0.9, with increments of 0.1}
- **Ooversampling Percentage**:{0.25,0.5,0.75,1,2,3,4}

The second method tested various combinations of oversampling and undersampling values for each class. We randomly selected combinations of the aforementioned values for testing.

To find the optimal combination of parameters and functions, we utilized two evaluation metrics across the Boosting algorithm and the full window evaluation: accuracy and a metric called TP, which is the sum of the individual accuracy of each class. The best two resampling strategies are in the Table A.1. The first strategy was denominated as Impsamp and the second as OveSamp.

Table A.1: Best resampling strategies

Function	Undersampling (%)			Oversampling (%)			
	LLS-PD	LHP-CEF	LHP-PB	FPO-PB	LHP-PBC	LLS-DEEP	PS-PB
Importance Sampling	0.3	0.3	0.3	4	4	4	4
UBL:Random OverClassification				1.25	4	1.5	1.75

Appendix B

Parameters Tested in Data Reconstruction Models

The following list describes the parameters tested in each machine learning algorithm in Section 4.1.1.2.

- **Decision Tree**: maximum depth: 10
- **Random Forest**: number of trees:{250, 500, 750}
- **Boosting**:
  - maximum depth:{2,10,20}
  - number of rounds:{50,75,100}
  - early stopping rounds: 20

The table B.1 presents the best parameters for Random Forest and Boosting in each time window. We calculated the maximum and minimum accuracy for each parameter over time to select the optimal parameters and then computed the difference between them. We ranked the parameters based on this difference to determine the best-performing parameter for each method and time window combination.

Table B.1: Best Parameters For Random Forest and Boosting for each window strategy

Method	Window	Parameters
Random Forest	Sliding	num.trees=750
	Growing	num.trees=250
	Full	num.trees=250
Boosting	Sliding	max depth= 10, nrounds= 100
	Growing	max depth= 10, nrounds= 75
	Full	max depth= 10, nrounds= 50

Appendix C

Hyperparameters for Length Prediction Algorithms

The following list describes the parameters tested in each machine learning algorithm in Section 4.2.3.

- **XGBoost**:
  - Number of estimators:{10,50,100,250}
  - Max depth:{3,5,10}
  - Learning rate:{0.001,0.01,0.1}
  - Objective function={"reg:squarederror"}
- **SVM**:
  - C:{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 100, 10000}
  - kernel:{"linear", "poly", "rbf"}
- **Neural Networks**:
  - Dense Layers: {(64), (100), (100, 64), (64, 64), (100, 100)}
  - Activation function:{"relu"}
  - Learning rate: {0.1, 0.01, 0.001}
  - Max epochs: {200}
  - Early Stopping:{True}
- **CNN**:
  - Convolution Layers: {1}
  - Number of filters: {(3,3)}
  - Number of filters :{8}

- Activation convolution layer:{"relu"}
- Max pooling:{(2,2)}
- Dense layers:{(64), (100), (100, 64), (64, 64), (100, 100)}
- Learning rate: {0.1, 0.01, 0.001}
- Max epochs: {200}
- Loss: {"mean_absolute_error"}
- Optimizer: {Adam from Tensorflow}
- Early Stopping: {from Tensorflow using Patient=2}

Bibliography

[1] Abdullah Al Muksit, Fakhrul Hasan, Md Fahad Hasan Bhuiyan Emon, Md Rakibul Haque, Arif Reza Anwary, and Swakkhar Shatabda. Yolo-fish: A robust fish detection model to detect fish in realistic underwater environment. *Ecological Informatics*, 72:101847, 2022.

[2] Amaya Álvarez-Ellacuría, Miquel Palmer, Ignacio A Catalán, and Jose-Luis Lisani. Image-based, unsupervised estimation of fish size from commercial landings using deep learning. *ICES Journal of Marine Science*, 77(4):1330–1339, 2020.

[3] Patrícia Amorim, Pedro Sousa, Ernesto Jardim, Manuela Azevedo, and Gui M Menezes. Length-frequency data approaches to evaluate snapper and grouper fisheries in the java sea, indonesia. *Fisheries Research*, 229:105576, 2020.

[4] Krishna Moorthy Babu, Daniel Bentall, David T Ashton, Morgan Puklowski, Warren Fantham, Harris T Lin, Nicholas PL Tuckey, Maren Wellenreuther, and Linley K Jesson. Computer vision in aquaculture: a case study of juvenile fish counting. *Journal of the Royal Society of New Zealand*, 53(1):52–68, 2023.

[5] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD Workshop*, 1994.

[6] Puja Bharati and Ankita Pramanik. Deep learning techniques—r-cnn to mask r-cnn: a survey. *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*, pages 657–668, 2020.

[7] Carl Boettiger, Duncan Temple Lang, and Peter Wainwright. rfishbase: exploring, manipulating and visualizing FishBase data from r. *Journal of Fish Biology*, nov 2012.

[8] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[9] Paula Branco, Rita P Ribeiro, and Luis Torgo. Ubl: an r package for utility-based learning. *arXiv preprint arXiv:1604.08079*, 2016.

[10] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[11] Xingyu Cai, Tingyang Xu, Jinfeng Yi, Junzhou Huang, and Sanguthevar Rajasekaran. Dtwnet: a dynamic time warping network. *Advances in neural information processing systems*, 32, 2019.

[12] N Carvalho. *Sea to shore: An economic evaluation of the Azorean Commercial Fisheries*. Phd thesis, University of the Azores, 2010.

[13] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[14] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4013–4022, 2018.

[15] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[16] Tianqi Chen, Tong He, Michael Benesty, and Vadim Khotilovich. Package 'xgboost'. *R version*, 90(1-66):40, 2019.

[17] Pau Climent-Pérez, Alejandro Galán-Cuenca, Nahuel Emiliano García-d'Urso, Marcelo Saval-Calvo, Jorge Azorin-Lopez, and Andres Fuster-Guillo. Automatic fish size estimation from uncalibrated fish market images using computer vision and deep learning. In *International workshop on soft computing models in industrial and environmental applications*, pages 319–329. Springer, 2022.

[18] Christopher Costello, Daniel Ovando, Ray Hilborn, Steven D Gaines, Olivier Deschenes, and Sarah E Lester. Status and solutions for the world's unassessed fisheries. *Science*, 338 (6106):517–520, 2012.

[19] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[20] TensorFlow Developers. Tensorflow. *Zenodo*, 2022.

[21] Caterina Dimitriadis, Ana Inés Borthagaray, Raul Vilela, Margarida Casadevall, and Alvar Carranza. A network approach to by-catch in a multi-species mediterranean small-scale fishery. *Fisheries Research*, 173:273–281, 2016.

[22] FAO. *The State of World Fisheries and Aquaculture 2024 – Blue Transformation in action*. FAO, 2024. ISBN: 978-92-5-138763-4.

[23] Leonardo N. Ferreira. From time series to networks in r with the ts2net package, 2022. doi:10.48550/ARXIV.2208.09660.

[24] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[25] Rainer Froese, Dirk Zeller, Kristin Kleisner, and Daniel Pauly. What catch data can tell us about the status of global fisheries. *Marine biology*, 159:1283–1292, 2012.

[26] Rainer Froese, Henning Winker, Gianpaolo Coro, Nazli Demirel, Athanassios C Tsikliras, Donna Dimarchopoulou, Giuseppe Scarcella, Martin Quaas, and Nele Matz-Lück. Status and rebuilding of european fisheries. *Marine Policy*, 93:159–170, 2018.

[27] Emma C Fuller, Jameal F Samhouri, Joshua S Stoll, Simon A Levin, and James R Watson. Characterizing fisheries connectivity in marine social–ecological systems. *ICES Journal of Marine Science*, 74(8):2087–2096, 2017.

[28] Rafael Garcia, Ricard Prados, Josep Quintana, Alexander Tempelaar, Nuno Gracias, Shale Rosen, Håvard Vågstøl, and Kristoffer Løvall. Automatic segmentation of fish using deep learning with application to fish size measurement. *ICES Journal of Marine Science*, 77(4): 1354–1366, 2020.

[29] Mark Gardener. *Beginning R: The statistical programming language*. John Wiley & Sons, 2012.

[30] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020.

[31] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[33] Ray Hilborn, Ricardo Oscar Amoroso, Christopher M Anderson, Julia K Baum, Trevor A Branch, Christopher Costello, Carryn L De Moor, Abdelmalek Faraj, Daniel Hively, Olaf P Jensen, et al. Effective fisheries management instrumental in improving fish stock status. *Proceedings of the National Academy of Sciences*, 117(4):2218–2224, 2020.

[34] Adrian Hordyk, Kotaro Ono, Keith Sainsbury, Neil Loneragan, and Jeremy Prince. Some explorations of the life history ratios to describe length composition, spawning-per-recruit, and the spawning potential ratio. *ICES Journal of Marine Science*, 72(1):204–216, 2015.

[35] Lorenzo Iafolla, Emiliano Fiorenza, Massimo Chiappini, Cosmo Carmisciano, and Valerio Antonio Iafolla. Sea wave data reconstruction using micro-seismic measurements and machine learning methods. *Frontiers in Marine Science*, 9:798167, 2022.

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[37] Mohamed S Kraiem, Fernando Sánchez-Hernández, and María N Moreno-García. Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. an approach based on association models. *Applied sciences*, 11(18):8546, 2021.

[38] Oliver Kramer and Oliver Kramer. Scikit-learn. *Machine learning for evolution strategies*, pages 45–53, 2016.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[40] Miron B Kursa, Aleksander Jankowski, and Witold R Rudnicki. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.

[41] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[42] Chunhong Liu, Zhiyong Wang, Yachao Li, Zhenzuo Zhang, Jiawei Li, Chen Xu, Rongxiang Du, Daoliang Li, and Qingling Duan. Research progress of computer vision technology in abnormal fish detection. *Aquacultural Engineering*, page 102350, 2023.

[43] Huan Liu, Zheng Liu, Shuo Liu, Yihao Liu, Junchi Bin, Fang Shi, and Haobin Dong. A nonlinear regression application via machine learning techniques for geomagnetic data reconstruction processing. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1): 128–140, 2018.

[44] LotacorSA. LOTAÇOR - Serviço de Lotas dos Açores, S.A. https://www.lotacor.pt/. Accessed on 20 June 2023.

[45] Catarina Maia, A Ferreira, Melinda Oroszlányová, Manuela Azevedo, P Ipma, P Lisbon, R Gaspar, C Silva, S Santos, and G Menezes. Fishmetrics: A new integrated solution for fisheries automatic and remote size data collection for fish stock assessment. In *Proc. 3rd Int. Conf. Elect., Electron., Eng. Trends, Commun., Optim. Sci.*, pages 1–7, 2016.

[46] Cyrus R Mehta, Nitin R Patel, and Pralay Senchaudhuri. Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association*, 83(404):999–1005, 1988.

[47] Pau Muñoz-Benavent, J Martínez-Peiró, Gabriela Andreu-García, Vicente Puig-Pons, V Espinosa, Isabel Pérez-Arjona, Fernando De la Gándara, and A Ortega. Impact evaluation of deep learning on image segmentation for automatic bluefin tuna sizing. *Aquacultural Engineering*, 99:102299, 2022.

[48] Keiko Nomura, Jameal F Samhouri, Andrew F Johnson, Alfredo Giron-Nava, and James R Watson. Fisheries connectivity measures of adaptive capacity in small-scale fisheries. *ICES Journal of Marine Science*, 79(2):519–531, 2022.

[49] S Nurdjaman, RA Wicaksono, AR Kartadikaria, and MR Abdullah. Data reconstruction of sea surface temperature in indonesia's fish management area 713 (ifma-713) using machine learning. In *IOP Conference Series: Earth and Environmental Science*, volume 1245, page 012037. IOP Publishing, 2023.

[50] Eko Prasetyo, Nanik Suciati, and Chastine Fatichah. A comparison of yolo and mask r-cnn for segmenting head and tail of fish. In *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–6, 2020. doi:10.1109/ICICoS51170.2020.9299024.

[51] Why Python. Python. *Python releases for windows*, 24, 2021.

[52] Joseph Redmon. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[53] Marie-Joëlle Rochet and Verena M Trenkel. Which community indicators can measure the impact of fishing? a review and proposals. *Canadian Journal of Fisheries and Aquatic Sciences*, 60(1):86–99, 2003.

[54] Natalie Rubio, Isha Datar, David Stachura, David Kaplan, and Kate Krueger. Cell-based fish: a novel approach to seafood production and an opportunity for cellular agriculture. *Frontiers in Sustainable Food Systems*, 3:43, 2019.

[55] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77:157–173, 2008.

[56] Régis Santos, Wendell Medeiros Leal, and Mário Rui Pinho. Stock assessment prioritization in the azores: Procedures, current challenges and recommendations. *Arquipelago. Life and Marine Sciences*, 37:45–64, 2020.

[57] Renuka Saravanan and Pothula Sujatha. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second international conference on intelligent computing and control systems (ICICCS)*, pages 945–949. IEEE, 2018.

[58] Jurgen Schmidhuber. Multi-column deep neural networks for image classification. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3642–3649, 2012.

[59] Vanessa Freitas Silva, Maria Eduarda Silva, Pedro Ribeiro, and Fernando Silva. Time series analysis via network science: Concepts and algorithms. *Data Mining and Knowledge Discovery*, 11(3), March 2021. doi:10.1002/widm.1404.

[60] Rajat Kumar Sinha, Ruchi Pandey, and Rohan Pattnaik. Deep learning for computer vision tasks: a review. *arXiv preprint arXiv:1804.03928*, 2018.

[61] Lev A Soinov, Maria A Krestyaninova, and Alvis Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome biology*, 4:1–10, 2003.

[62] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[63] Shan Suthaharan and Shan Suthaharan. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235, 2016.

[64] Russ Tedrake. *Robotic Manipulation*. Course Notes for MIT 6.421, 2024.

[65] Pauli Virtanen, Ralf Gommers, Evgeni Burovski, Travis E Oliphant, Warren Weckesser, David Cournapeau, Pearu Peterson, Tyler Reddy, Matt Haberland, Josh Wilson, et al. scipy/scipy: Scipy 1.6. 0. *Zenodo*, 2021.

[66] Wenhao Xie, Gongqian Liang, Zhonghui Dong, Baoyu Tan, and Baosheng Zhang. An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data. *Mathematical Problems in Engineering*, 2019(1):3526539, 2019.

[67] Yaozhen Yu, Hao Zhang, and Fei Yuan. Key point detection method for fish size measurement based on deep learning. *IET Image Processing*, 17(14):4142–4158, 2023.

[68] Jiushuang Zhang and Yong Wang. A new workflow for instance segmentation of fish with yolo. *Journal of Marine Science and Engineering*, 12(6):1010, 2024.

[69] Shili Zhao, Song Zhang, Jincun Liu, He Wang, Jia Zhu, Daoliang Li, and Ran Zhao. Application of machine learning in intelligent fish aquaculture: A review. *Aquaculture*, 540: 736724, 2021.

[70] Chong Zhou. *Yolact++ Better Real-Time Instance Segmentation*. University of California, Davis, 2020.