# Integrated approaches for monitoring sharks: Leveraging machine learning, big data, and molecular biology

Jeremy F. Jenrette

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Fish and Wildlife Conservation

Francesco Ferretti, Chair
Edward Fox
Eric Hallerman
Leah Johnson

September 16, 2025
Blacksburg, Virginia

Keywords: Big data, Environmental DNA, Machine Learning, Sharks

# Integrated approaches for monitoring sharks: Leveraging machine learning, big data, and molecular biology

Jeremy F. Jenrette

(ABSTRACT)

Sharks are ecologically important predators facing severe global declines, yet conservation and management are hindered by data deficiencies in taxonomy, distribution, and abundance. In this dissertation, I develop and integrate complementary technological approaches: machine learning, big data workflows, and molecular techniques—to expand scalable, non-invasive monitoring of sharks with programmatic and practical field methodologies. First, I constructed the largest global shark image dataset to date and developed the Shark Detector, a pipeline combining object detection and hierarchical classification. This system automatically locates, identifies, and classifies sharks in heterogeneous media, achieving >90% recall for detection and up to 92% species-level classification accuracy across 80 species, outperforming existing biodiversity classifiers. Second, we refined these methods for ecological survey applications by packaging the models into `sharkDetectoR` (R package) and SharkByte (desktop application), enabling accessible, semi-automatic processing of baited remote underwater videos (BRUVs). These tools reduced annotation effort by up to 95% while preserving high taxonomic resolution, and demonstrated iterative improvement through survey-specific data boosting. Third, I designed scalable pipelines to mine and filter >5 million social network (Instagram, Flickr) and open source (iNaturalist and Global Biodiversity Information Facility) posts and >600k opportunistic shark observations. By pairing automated classification with effort-standardized statistical models, I derived species-specific abundance indices that revealed regionally consistent population trends: increasing trajectories for coastal taxa

in the Bahamas, and recent declines of reef-associated sharks in the Hawaiian Islands. Finally, I piloted molecular monitoring of critically endangered white sharks (*Carcharodon carcharias*) in the Mediterranean Sea using complimentary Environmental DNA detection and validation workflows. I collected 204 samples across the Sicilian Channel, Adriatic and Ligurian Seas, and detected white sharks at four stations. Detections were confirmed in the lab. Particle simulations identified the detected individuals as nearby for the purpose of tracking them in the field. A preliminary multi-species assay detected 12 elasmobranch species. These workflows provided novel spatiotemporal insights into white shark (and other elasmobranch) occurrence in hypothesized hotspots. Together, these chapters demonstrate how integrated computational and molecular approaches can overcome data limitations, provide reproducible ecological indices, and inform conservation of threatened shark populations in data-poor regions.

# Integrated approaches for monitoring sharks: Leveraging machine learning, big data, and molecular biology

Jeremy F. Jenrette

(GENERAL AUDIENCE ABSTRACT)

Sharks are vital to healthy oceans but remain among the most threatened and data-poor groups of animals, largely because they are difficult to monitor. This dissertation develops new ways to study sharks using artificial intelligence, online citizen science data, and Environmental DNA. I built the largest collection of shark images ever assembled and trained computer models to automatically find and identify species in photos and videos, making surveys faster and more accurate. I created tools that allow researchers and citizen scientists to process underwater footage on their own computers, greatly reducing the time required to review hours of video. By analyzing millions of shark images shared on Social Networks and biodiversity websites, I uncovered patterns of abundance that reflect real population trends. I identified coastal shark numbers rising in the Bahamas but declining around the Hawaiian Islands. Finally, I tested Environmental DNA techniques to detect critically endangered white sharks (*Carcharodon carcharias*) in the Mediterranean Sea, successfully finding their genetic traces in the Sicilian Channel, Adriatic and Ligurian Seas. Together, these approaches show how combining big data, machine learning, and molecular methods can fill major knowledge gaps and provide new tools to protect sharks worldwide.

# Dedication

*I dedicate this dissertation to my parents, Bruce and Jennifer Jenrette.*

# Acknowledgments

I acknowledge the following funding sources: The Explorers Club, The Discovery Channel, the Center for Coastal Studies, and the Acorn Alcinda Foundation. I thank the National Geographic Society (NGS) for funding the Meridian Project through the Large Marine Vertebrates Research Institute Philippines, and to the NGS Expedition Donors for sponsoring the expedition. I thank the Ocean Exploration Trust management and crew for providing the opportunity to work aboard *E/V Nautilus* and support the team throughout the entire expedition, as well as for coordinating and facilitating the necessary permits to be able to work in the Hawaiian Islands. I extend my gratitude to the Hawaii-based dive operators Extended Horizons and Liquid Cosmos Divers for their time, expertise, and resources in assisting the field team accomplish the field operations. I thank Filippo Varini and the other developers of *SharkTrack* for making their detection model open source. I thank the faculty and staff at the Virginia Tech Genomics Sequencing Center for their lab space, mentorship, and services. For their assistance in data collection, validation, and backend administration, I thank Virginia Tech undergraduate and graduate students Tuan Tran, Mia Hagood, Patrick Warner, Aakash Divakar, Omar Kalbouneh, and Aman Kothari from the Computer Science Department, and Lauren Morris from the Fish and Wildlife Department. I thank graduate students Amritha Subramanian and Aseem Sangwan from the Computer Science Department. I acknowledge the data-validation efforts performed by high school student Molly Fuchs as well. I thank the Virginia Tech Computer Science Department and Rob Hunter for administrating various virtual machines. And I thank my committee, Francesco, Edward, Eric, and Leah for their continued support and mentorship.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**Adagrad** Adaptive Gradient Algorithm. xxiv, 18

**Adam** Adaptive Moment Estimation. xxiv, 14, 45

**AI** Artificial Intelligence. xxiv, 3, 33, 39, 41–43, 52, 62, 122–124, 127, 176

**API** Application Programming Interface. xiv, xxiv, 39, 46, 51, 60, 66, 70, 71, 80, 123, 173

**AUC** Area Under the Curve. xii, xxiv, 20

**BRUV** Baited Remote Underwater Video. xiv, xxi–xxiv, 2, 6, 8, 9, 11, 12, 18, 26, 35, 36, 39–41, 43–45, 49, 50, 52, 53, 56, 58–62, 64, 67, 78, 91, 93, 94, 96, 104, 120, 122–124, 126, 127, 184–186

**CNN** Convolutional Neural Network. xiv, xxiv, 4, 8, 14, 15, 36, 41, 44, 45, 48, 60, 177

**COCO** Common Objects in Context. xxiv, 12, 17

**CPUE** Catch Per Unit Effort. xxiii, xxiv, 64, 78, 91, 93–96

**CYTB** Cytochrome B. xviii, xxiv, 107, 109, 110, 112

**DenseNet201** Densely Connected Convolutional Network (201 layers). xxiv, 18

**DNA** Deoxyribonucleic Acid. xxiv

**DwC** Darwin Core. xxiv, 72

**eDNA** Environmental DNA. xvii–xix, xxiv, 3–6, 40, 49, 52, 99–122, 124, 125, 127, 128, 130, 167

**EEZ** Exclusive Economic Zone. xxiv, 185

**$F_1$** $F_1$ Score. xiv, xx, xxiii, xxiv, 21, 23, 31, 50, 54, 56–58, 178, 180, 181

**Faster-R-CNN** Faster Region-based Convolutional Neural Network. xxiv, 12, 31

**FN** False Negative. xxiv, 23, 54, 61, 103

**FNR** False Negative Rate. xxiv, 23

**FP** False Positive. xviii, xxiv, 19, 23, 54, 101, 103, 109, 111, 113, 120, 124

**FPR** False Positive Rate. xxiv, 23

**GAM** Generalized Additive Model. xxiv

**GBIF** Global Biodiversity Information Facility. xv, xxiv, 5, 64–67, 72, 73, 79, 84, 94, 123

**GFW** Global Fishing Watch. xxiv

**GLM** Generalized Linear Model. xxiv, 69, 84

**GPU** Graphics Processing Unit. xxiv, 46, 49, 60

**GSC** Genus Specific Classifier. xiii, xxii, xxiv, 16, 17, 21, 22, 27–31

**GUI** Graphical User Interface. xv, xxi, xxiv, 35, 47, 55, 123, 126, 185, 186

**IG** Instagram. xv, xix, xxii, xxiv, 5, 8, 12, 14, 16–18, 21, 23, 25, 32, 33, 64–67, 70–73, 79–81, 83, 94, 122, 123, 168, 169, 187

**iNat** iNaturalist. xv, xvi, xxiii, xxiv, 5, 17, 21, 29, 32, 64–67, 72–74, 79, 81, 82, 84, 87–95, 122–124

**IUCN** International Union for Conservation of Nature. xx, xxiv, 1, 9, 47, 77, 100, 121, 126, 130, 174, 175

**MaxN** Maximum Number. xxiii, xxiv, 56–59, 61, 78, 91, 93

**MCMC** Markov Chain Monte Carlo. xxiv

**MEDITS** Mediterranean International Trawl Surveys. xxiv, 119, 120

**MHI** Main Hawaiian Island. xiv, xvi, xxiii, xxiv, 43, 49, 50, 52, 53, 57–59, 61, 64, 75, 82, 87, 88, 90, 92, 94, 95, 170, 175, 184

**ML** machine learning. xxiv, 8, 10

**MPA** Marine Protected Area. xxiv, 126, 127

**NB** Negative Binomial. xxiv

**NGO** Non-Governmental Organization. xxiv

**NLP** Natural Language Processing. xxiv

**NMFS** National Marine Fisheries Service. xxiv

**NOAA** National Oceanic and Atmospheric Administration. xxiv, 126, 129, 130

**NWHI** Northwestern Hawaiian Islands. xxiv

**PCR** Polymerase Chain Reaction. xvii, xviii, xxiv, 105, 107, 108, 110, 111, 113

**psql** PostgreSQL interactive terminal. xxiv

**PyQt5** Python Qt version 5 (comprehensive Python bindings for the Qt framework). xxiv, 47

**ReLU** Rectified Linear Unit. xxiv, 14, 18

**ROC** Receiver Operating Characteristic. xii, xxiv, 20

**ROV** Remotely Operated Underwater Vehicle. xxiv, 11, 40

**SC** Shark Classifier. xii–xiv, xix, xxii, xxiv, 11, 13, 16–19, 24–26, 30, 31, 35, 45, 46, 48, 56, 60, 168, 169, 171–173, 176

**SD** Shark Detector. xii–xiv, xx–xxii, xxiv, 4–6, 8, 11, 13, 17–19, 21, 22, 26, 29, 32–37, 39, 42, 43, 45, 47, 48, 50, 61, 62, 71, 72, 74, 80, 81, 85, 94, 122, 123, 127–129, 168, 170, 175, 177–183, 185, 186

**SEDAR** Southeast Data, Assessment, and Review. xxiv

**SfM** Structure-from-Motion. xxiv

**SI** Shark Identifier. xii, xiii, xix, xxii, xxiv, 11–14, 19, 20, 23, 24, 26, 31, 35, 168, 169

**SL** Shark Locator. xii, xxii, xxiv, 11–13, 15, 16, 18, 19, 21, 26, 31, 35

**SN** social network. xxii, xxiv, 2, 4–6, 10, 11, 17, 18, 33, 36, 42, 44, 64–67, 72, 77, 79, 80, 90, 92, 123, 124

**SNP** Single Nucleotide Polymorphism. xviii, xxiv, 109, 110

**SP** sharkPulse. xxiv, 17

**SPUE** Sightings per Unit Effort. xvi, xxiii, xxiv, 34, 64, 67, 69, 77, 79, 84–96, 126

**SQL** Structured Query Language. xxiv

**SSCg** Genus-specific Species Classifier. xiii, xxii, xxiv, 16, 17, 21, 22, 26, 28, 31

**SST** Sea Surface Temperature. xxiv

**TN** True Negative. xxiv, 23, 54

**TP** True Positive. xxiv, 23, 54

**VGG16** Visual Geometry Group 16-layer network. xxiv, 14, 31

**VM** virtual machine. xxiv

**VT-ARC** Virginia Tech Advanced Research Computing. xxiv, 45

**VT-GSC** Virginia Tech Genomics Sequencing Center. xvii, xxiv, 105–107, 111, 127

**YOLO** You Only Look Once (object detection algorithm). xiv, xxiv, 39, 44, 48, 60

**YT** YouTube. xx, xxii, xxiv, 17–19, 26, 171, 173–176

# Chapter 1

# Introduction

## 1.1 The Problem: Deficiency of Shark Observational Data

Sharks are among the oldest and most evolutionarily resilient vertebrates on Earth, having persisted for over 400 million years and survived multiple mass extinctions [35]. Despite this evolutionary success, modern shark populations are declining at alarming rates. Since 1970, global abundance of oceanic sharks and rays has decreased by more than 70%, driven primarily by overfishing, bycatch, and habitat degradation [117]. This crisis has been compounded by the effects of climate change and anthropogenic effects [44, 68].

Sharks are characterized by slow growth, late sexual maturity, and low fecundity, rendering them particularly vulnerable to overexploitation [146]. Yet, while their ecological importance as apex and mesopredators is well established, fundamental knowledge of population size, structure, and distribution remains incomplete for most species [31]. Nearly half of all assessed shark species are listed as threatened or data deficient by the IUCN [44]. This lack of standardized, high-resolution data limits the ability to assess trends, predict extinction risk, and develop effective management strategies.

Conventional monitoring approaches—such as long-term fisheries surveys, observer programs, and tagging studies—are costly, geographically restricted, and logistically challenging

to maintain, particularly in developing regions [15, 54, 134, 135]. As a result, shark conservation measures are often delayed or unsuccessful, often relying on anecdotal evidence or limited catch statistics [14, 40]. In the absence of supplemental data streams, population declines continue largely undetected, and conservation is insufficient to reverse them [40, 117]. With glaring ecological data gaps describing imperiled populations, we need to integrate underutilized data sources. This dissertation addresses how to harness those sources for bridging knowledge disparity.

## 1.2 Emerging Technologies

The big data revolution has transformed how scientists acquire and interpret ecological information. Vast amounts of data are continuously produced by humans in the form of images and videos that capture wildlife encounters and ecological activity. SNs, open biodiversity platforms, and citizen-science repositories now reveal unprecedented digital records of wildlife occurrences across space and time [64, 108]. These data streams are largely unexplored for inferring ecological patterns, but with the growth of big data and machine learning, we have the tools to systematically extract and transform digital information into ecologically relevant observations that expand the temporal and spatial reach of biodiversity monitoring. At the same time, similar computational approaches are revolutionizing independent underwater surveys such as BRUV, where automated detection and classification can mitigate the need for exhaustive manual review and speed up quantitative analysis of marine wildlife. The evolution of these tools toward accessible and intuitive interfaces ensures that scientists, managers, and citizens alike can explore, filter, and interpret ecological data with unprecedented ease and transparency [168].

In parallel, molecular approaches have advanced rapidly to complement visual and dig-

ital monitoring. eDNA techniques detect genetic material shed by organisms into their surroundings, providing highly sensitive and non-invasive indications of presence that capture both rare and cryptic taxa [110, 158]. When coupled with metabarcoding and next-generation sequencing, eDNA enables community-level assessments of biodiversity that are often unattainable through visual surveys alone.

The Mediterranean white shark is an illustrative example of a population that has been pushed to the edge of extinction in an exploited and data-poor region [54, 111, 118]. Visual observations are incredibly rare. Thus, the ability to detect their presence with eDNA is a distinct advantage for identifying their last strongholds and guiding targeted conservation efforts [54].

The intersection of big data, Artificial Intelligence (AI), and molecular forensics holds particular promise for shark conservation. Online imagery can provide indicators of relative abundance, behavior, and spatial distribution [52, 111, 150]. Molecular assays can reveal cryptic or endangered populations that evade visual detection [36]. When demonstrated with rigorous analytical frameworks, these approaches can expand the geographic and temporal coverage of monitoring while maintaining transparency and reproducibility [161, 166, 169]. Harnessing these underutilized data streams complements traditional science by integrating automation into data collection, greatly increasing the speed and scope of ecological monitoring.

The overarching goal of this dissertation is to establish an innovative framework for overcoming data deficiency in shark conservation. Thus, this dissertation pursues three interrelated objectives:

1. Develop and validate deep learning models that automatically detect and classify visual media of sharks to the species level.

2. Construct a big data workflow that aggregates, cleans, and models shark observations from major SNs and online platforms to estimate relative abundance and assess temporal trends across case-study regions.

3. Demonstrate a robust white shark-specific eDNA assay, coupled with oceanographic modeling and citizen science sampling, to map the occurrence of Mediterranean shark populations.

Together, my approaches demonstrate workflows for classifying and monitoring sharks across the digital and molecular domains. My objective was developing and refining methodological approaches that are either emerging or unconventional for generating conservation information.

## 1.3 Chapter Outlines

To harness these emerging technologies, this dissertation builds upon the foundations of *sharkPulse*—a global cyberinfrastructure and research initiative designed to transform digital observations into ecological insight [52]. The platform currently hosts over 300k shark observations representing 309 species, aggregated from SNs, open biodiversity repositories, and user submissions. These data supply the training and validation structure of a package of programs, the SD, forms the backbone of the programmatic and analytical approaches of this dissertation.

Modern deep learning approaches power image recognition and enable automated classification of organisms with accuracy approaching that of human experts [116, 131]. CNNs now underpin many biodiversity frameworks that rely on computer vision. Yet, despite this progress, sharks remain underrepresented in open models, exhibiting substantial visual

variability across life stages, lighting conditions, and viewing angles. Establishing robust, generalizable taxonomic identifiers for sharks therefore demands both model innovation and carefully annotated training datasets [169]. Within this context, Chapters 2 and 3 develop the SD as an ensemble of object-detection and classification models that automate species-level identification from heterogeneous imagery and underwater footage, made more accessible through an R package `sharkDetectoR` [75] and a graphical interface. Over the course of this study, the SD grows smarter as new data is ingested and incorporated into the training architecture. New versions represent increased capability of the SD to classify a larger range of taxa at a higher performance.

SNs and open biodiversity platforms now generate massive quantities of geo-tagged images that record human–wildlife encounters at global scales [64, 108]. However, these platforms differ in their data structure and reliability. IG and Facebook, Flickr, iNat, and GBIF exemplify the diversity of publicly accessible ecological observations. However, the heterogeneity of these data sources presents both a challenge and an opportunity: while the raw observations are variably tagged or incomplete, their combined volume and coverage far exceed those of traditional surveys. Chapter 4 addresses whether it is feasible to crowdsource and interpret ecological patterns from such repositories by developing a semi-automated workflow for collecting, cleaning, and modeling shark observations. By estimating relative abundance and temporal trends through negative binomial models adjusted for user activity, the framework evaluates how SN data can supplement conventional indices of relative abundance.

The application of eDNA in the open ocean remains technically challenging: concentrations of target DNA are often orders of magnitude lower than in coastal or freshwater systems, and degradation, transport, and shedding dynamics vary with environment and animal physiology. Reliable detection thus depends on rigorous sampling design, contamination control, and laboratory precision. Chapter 5 integrates these methodological safeguards into a tar-

geted 4-year long eDNA survey for detecting the critically endangered Mediterranean white shark. In these efforts, I explore particle tracking simulations to estimate real-time spatial presence for the purpose of finding and tagging individuals. I also develop citizen science water-sampling kits for establishing a network of training samplers and increasing sampling distribution and intensity. And lastly, I explore how to estimate broader biodiversity by incorporating elasmobranch-specific primers in conventional metabarcoding approaches.

Overall, I automate image classification, SN analytics, and molecular detection. These chapters represent complementary pathways toward resolving the data-deficiency barrier that impedes shark conservation. The integration of these technologies within the *sharkPulse* framework establishes a transparent and reproducible system for scaling classification of sharks and supplementing traditional monitoring techniques. In Chapter 6, I develop conclusions on the effectiveness and practicality of Chapters 2, 3, 4, 5 to boost population-level monitoring, and their conservation relevance. To address under-reporting of shark bycatch at the commercial level, I provide future recommendations for applying the SD to electronic monitoring scenarios. I suggest methods to boost eDNA analyses with reference databases, and how to incorporate tagging and BRUVs for validating species detections, contextualizing habitat use and movement patterns across spatial and temporal scales. To extend this work, I provide the standing software repositories and URL links in Appendix A.

# Chapter 2

# Shark Detection and Classification

Published as J. Jenrette, Z. Liu, P. Chimote, T. Hastie, E. Fox, and F. Ferretti. Shark detection and classification with machine learning. *Ecological Informatics*, 69:101673, 2022. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2022.101673. URL https://www.sciencedirect.com/science/article/pii/S1574954122001236

The material presented here is adapted from the published article with updates, clarifications, and supplementary analyses.

# Abstract

Suitable shark conservation depends on well-informed population assessments. Direct methods such as scientific surveys and fisheries monitoring are adequate for defining population statuses, but species-specific indices of abundance and distribution coming from these sources are rare for most shark species. We can rapidly fill these information gaps by boosting media-based remote monitoring efforts with machine learning (ML) and automation. We created a database of 53,345 shark images covering 219 species of sharks, and packaged object detection and image classification models into a *Shark Detector* (Shark Detector (SD)) bundle. The SD recognizes and classifies sharks from videos and images using transfer learning and Convolutional Neural Networks (CNNs). We applied these models to common data-generation approaches of sharks: collecting occurrence records from photographs taken by the public or citizen scientists, processing baited remote underwater video Baited Remote Underwater Video (BRUV) footage and online videos, and data-mining Instagram (IG). We examined the accuracy of each model and tested genus and species prediction correctness as a result of training data quantity. The SD can classify 47 species pertaining to 26 genera. It sorted heterogeneous datasets of images sourced from IG with 91% accuracy and classified species with 70% accuracy. It located sharks in BRUV footage and YouTube videos with 89% accuracy, and classified located subjects to the species level with 69% accuracy. All data-generation methods were processed without manual interaction. As media-based remote monitoring appears to dominate methods for observing sharks in nature, we developed an open-source SD to facilitate common identification applications. Prediction accuracy of the software pipeline increases as more images are added to the training dataset. We provide public access to the software on our GitHub page.

## 2.1 Introduction

Sharks are excellent indicators of ocean environmental health: however, they are constantly challenged by growing fishing pressures as well as poor management and conservation stemming from data paucity, insufficient taxonomic knowledge, and underdeveloped monitoring methods [82]. Observation data of sharks via surveying and fisheries monitoring are often extremely costly or difficult to collect, a challenge exacerbated for species with larger home ranges [13]. Furthermore, classification of sharks is still debated for many species [134]. The combination of observed global declines and increasing data resolution has resulted in the number of International Union for Conservation of Nature (IUCN)-listed threatened species doubling since 2014 [117]. Sharks remain an extremely data-deficient group of marine animals, and these information gaps contribute to the lack of abundance and distribution indices as well as taxonomic precision needed to properly assess population statistics [52, 82].

Image-based biomonitoring is a transformative alternative to expensive and invasive direct observation methods in ecological surveys of marine and terrestrial environments [139, 169, 171]. With significant advancements in Baited Remote Underwater Video (BRUV) systems, motion-activated camera traps, and crowdsourced citizen science media, ecological information is being produced at an unprecedented rate [60]. Importantly, remote monitoring methods generate visual media that can help fill shark information gaps. These methods are non-invasive and useful for minimizing sampling effort: however, they produce large quantities of media to post-process for species identification and analyses, including removing irrelevant images [151]. Studies such as Tabak et al. [152] and Malde et al. [101] stress the importance of using deep learning programs to filter unrelated content and facilitate rapid sampling.

Deep learning algorithms are highly flexible and well suited for approaching many of these tasks [95, 101, 139]. They have been used to estimate fish sizes from images [47, 59], identify discarded and processed fish [56], and classify acoustic and movement data [20, 49, 83]. However, machine-learned detection and image classification of shark species are seldom studied due to insufficient training data [52]. Video and photographic documentation of sharks are rarely obtained from commercial fisheries: such images are more reliably sourced from tourists, social network (SN)s, and underwater photographers [153]. Consequently, there are few studies that have curated a training dataset of shark images. *iSharkFin* is a recognition system that can identify 39 shark species from pictures of dorsal fins [9]: however, it does not classify whole-body images because it focuses primarily on tracing illegal fin trading and requires users to manually select features and input points that describe fin shape. As a step forward, *Seek* is a generalist image classifier that leverages the iNat2017 archive of 859,000 images to detect and classify over 5,000 organisms, including shark species [69]. This app advances animal classification, but because of its large scope, it remains inaccurate for classifying the 509 species of living sharks.

Because many shark species are both morphologically diverse and data-poor, classifying them is not straightforward for machine learning (ML). Here, we approach this challenge by first constructing the largest training dataset of shark images. Second, we combine object detection and hierarchical classification methods for images and videos to facilitate the creation of biologically relevant data on sharks. Assembling and annotating large amounts of data-mined and user-uploaded media for conservation use is an emerging approach [108, 149]. Fish species classification with ML algorithms has only begun developing within the last two decades [139]. As a result of interacting with big data and citizen scientists, we have created the largest and most diverse archive of shark images. Few studies have combined object detection with classification to increase shark taxonomic accuracy [9, 69]. Our objective was

to automatically detect and classify, to the species level, any image with perceptible shark features. Because our methods build upon standard recognition and data-mining approaches, we can generate, detect, and classify shark-sourced visual media. We can efficiently post-process BRUV footage, camera trap images, Remotely Operated Underwater Vehicle (ROV) footage, and shared social media by automatically removing irrelevant content and classifying shark species.

## 2.2 Methods

Our shark detection and classification pipeline is composed of several steps and three main components (Figure 2.1): (1) an object detection model called the *Shark Locator* (Shark Locator (SL)), which locates one or several shark subjects in images and draws bounding boxes around them: (2) a binary sorting model called *Shark Identifier* (Shark Identifier (SI)), which sorts images of sharks from a pool of heterogeneous images: and (3) multiclass models called *Shark Classifiers* (Shark Classifier (SC)), which classify shark images to the genus and species levels. Combining these three modeling components, we developed a shark identification and classification pipeline called *Shark Detector* (Shark Detector (SD)), which can ingest any media containing shark subjects, locate and sort subjects according to relevance, and classify the sharks to the species level. Shark training images for developing these models were mainly sourced from sharkPulse—a crowd-sourcing platform that mines and aggregates shark media from SNs, citizen science projects, user submissions, and other electronic archives [52].

### 2.2.1 Shark Locator: Object Detection

The identification pipeline starts with the SL. This model is primarily used to inflate an initial training dataset by cropping one or multiple shark subjects from images, thereby creating new images. It locates shark subjects in videos (e.g., BRUV footage) and extracts frames with shark subjects. This process had the dual objective of removing irrelevant subjects or noisy backgrounds that challenged the training process, and boosting the training dataset by splitting images with multiple shark subjects into multiple distinct shark training images. Cropping shark features from images and video frames provided better training quality.

To build the SL, we sourced TensorFlow's Model Garden [175] and used a Faster Region-based Convolutional Neural Network (Faster-R-CNN) algorithm [123]. The model was trained with the Common Objects in Context (COCO) dataset (consisting of 236 shark images) to detect and draw boxes around sharks [99]. Faster-R-CNN can detect more than one object within a frame, allowing multiple boxes to be drawn [123]. To reduce processing time, we set a limit of 10 boxes that could be drawn within a single frame. The SL boosted our classification dataset from 24,546 images to 53,345 images.

### 2.2.2 Shark Identifier: Binary Model

Second, we developed a binary sorting model. The SI identifies shark vs. non-shark subjects in images and is used to filter out non-shark images before the remaining images are taxonomically classified. We sourced 53,345 shark images from Instagram (IG) and sharkPulse, and additionally, we sourced 50,260 non-shark images from IG (Table 2.2). First, we constructed the SI to learn key shark features from training images by optimization of the binary cross-entropy loss function [95]. We incorporated a pre-trained model to reduce

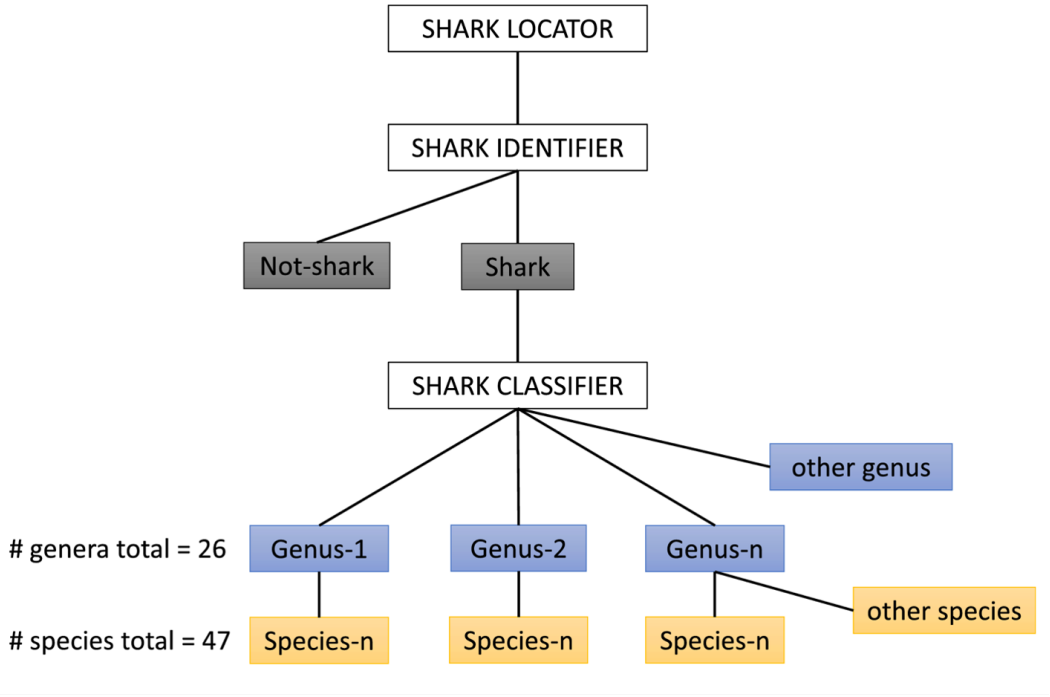


Figure 2.1: The SD system is composed of object detection and classification packages that work best in a stepwise procedure. Additionally, by detecting shark subjects, the SL synthetically supplements the sharkPulse archive with cropped shark images available to SI and SC models as new training data. Videos are processed in the order of locating, identifying, and then classifying. Heterogeneous data-mined datasets are processed in the order of identifying and then classifying.

the number of training steps (transfer learning). The SI was pre-trained with the Visual Geometry Group 16-layer network (VGG16) network, which was trained on 1.28 million images with 1,000 categories and achieves 92.7% test accuracy on the ImageNet dataset [140]. Convolutional Neural Networks (CNNs) perform best when the categories of interest are well represented and balanced [100]. Non-shark images were sourced entirely from IG. We resized images to $150 \times 150$ pixels to reduce memory consumption.

To increase training accuracy, we used positional image augmentation techniques, i.e., we artificially augmented images with transformations such as width and height shifting, shearing, zooming, and rotations [152, 154]. Then, we constructed convolutional-pooling layers, which act as checkpoints for summarizing features the model has learned [95]. When shark features are learned from trained images, the CNN generates parameters called weights. Weights were first initialized when we pre-trained networks on the ImageNet dataset. We froze the bottom pre-trained layers to prevent weights from being modified while we trained the top layers for shark features. VGG16 contains 16 pre-trained layers, and we added four layers to train for shark features. We trained the CNN to accept augmented and regular training images as raw pixels and gradually learn output predictions as they passed through convolutional-pooling layers. To facilitate adaptive learning, we adjusted the algorithm's learning rate to $5 \times 10^{-4}$ with the Adaptive Moment Estimation (Adam) optimizer [89]. To avoid vanishing gradients and improve training speed, we incorporated Rectified Linear Unit (ReLU) activation into the CNN's fully connected layers [113]. The output layer was composed of two neurons for classification, corresponding to the number of classes being trained: shark and non-shark. These neurons were normalized with a softmax activation function [21]. We trained the model with 90% of the training set and validated with the remaining images over 10 epochs. One epoch represents one full cycle where the algorithm has processed the entire training dataset. To prevent the model from overfitting, we incorporated

dropout and regularization parameters. Dropout effectively removes a percentage of neurons from the model that have learned features, and helps the model generalize when predicting new images. We set dropout to 30%. Image augmentation and early stopping are forms of regularization used to minimize validation error [144]. We incorporated early stopping of training when test accuracy decreased for three consecutive epochs. Then we measured the curve in Figure 2.3. We built the CNN in Python using the Keras and TensorFlow packages [1, 30].

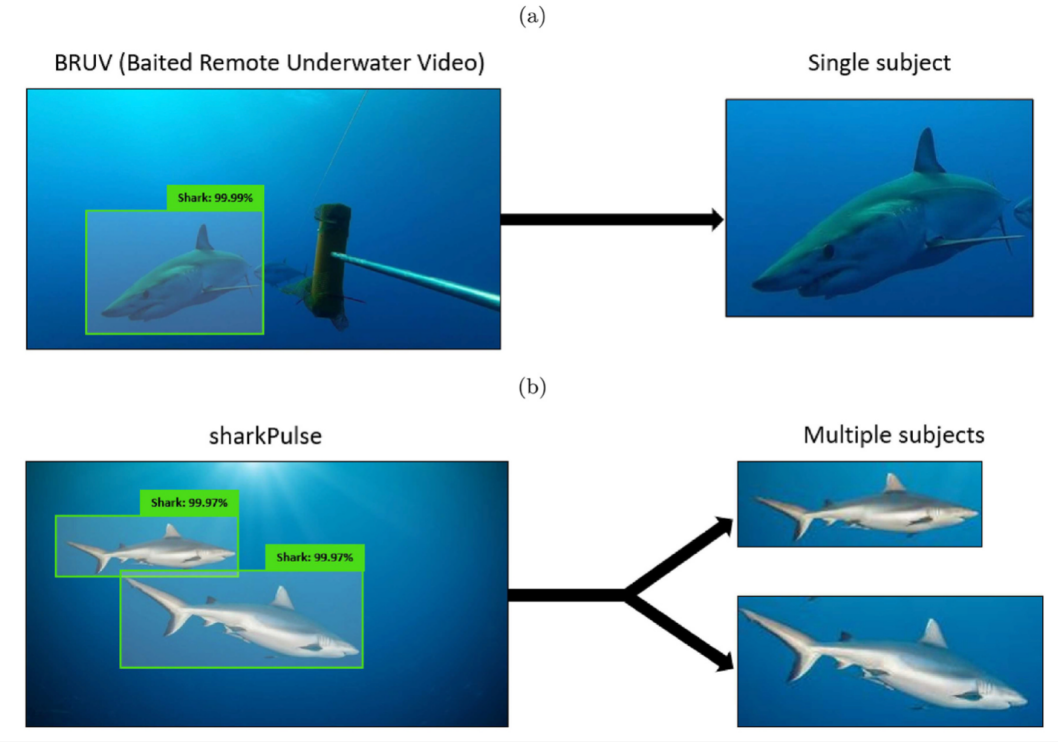


Figure 2.2: The SL object detection model draws boxes corresponding to confidence levels of shark presence. (a) A juvenile shortfin mako is detected and a single auto-cropped image is processed, removing irrelevant objects such as the bait canister and bluefin tuna. (b) Multiple *Carcharhinidae* species are detected and two images are cropped from a single image.

### 2.2.3 Shark Classifier: Genus and Species classification

We developed the SC as a hierarchical classification framework for classifying the identified shark images taxonomically. We trained one genus-specific model and a series of local species-specific models—one for each genus (Figure 2.1). The SC ingests the filtered shark images and classifies them at the genus level with the Genus Specific Classifier (GSC). Then, depending on the genus, a Genus-specific Species Classifier (SSCg) predicts the most likely species.

We trained the SC with the sharkPulse database, images cropped with the SL, and IG images. In total, the SC contains 74 genera and 219 species of sharks with an average of 167 images per species (Table 2.1). The GSC was trained with 36,722 images, and the SSCg was trained with 19,243 images. We evaluated the recall of a genus class vs. its training data quantity in Figure 2.4a, which revealed an average of $433 \pm 47$ images were needed to produce $\geq 50\%$ recall. Recall measures the proportion of shark images that were correctly classified. Most genera ($> 64\%$) did not contain this many images to produce adequate training quality. Once a genus was classified, we looked at the same relationship for species classes in Figure 2.4b and discovered an average of $161 \pm 41$ images were needed to produce $> 50\%$ recall. We used these averages as training data quantity thresholds for the SC (see Figures 2.4c and 2.4d).

Next, to better understand why models confused classes with each other, we examined two metrics. We used Pielou's evenness index, usually employed to assess whether and to what extent species' abundances are uniform in a community, to quantify how balanced training datasets were [119]. Then we evaluated the difference in morphology by calculating the Euclidean distance between species. This was done by collecting a common set of morphometric measurements for 124 species that were available in the `rfishbase` R pack-

age and that represented our dataset [17, 116]. We used total length, standard length, fork length, and head length. Then we calculated the average for each measurement to create a morphometric centroid of all species. We compared each species to this centroid to assess morphological homogeneity.

| Model | # of models | Training images | Test images | Training source |
|---|---|---|---|---|
| Shark Locator | 1 | 236 | — | COCO dataset |
| Shark Identifier | 1 | 93,244 | 10,361 | sharkPulse (SP), Flickr, iNaturalist (iNat), IG, YouTube (YT) |
| Genus-specific Classifier | 1 | 33,050 | 3,672 | SP, Flickr, iNat, IG, YT |
| Species-specific Classifier | 18 | 17,319 | 1,924 | SP, Flickr, iNat, IG, YT |

Table 2.1: SD packages trained with images sourced from various SNs and online archives.

For the GSC, we trained 36,722 images across 26 genus classes that met the training data threshold of 433 images (Table 2.1). We trained a 27th class with 2,593 images to represent the $> 64\%$ of genera that did not meet the training data threshold. This class was labeled "other genus." Since there were 18 shark genera containing two or more species, we developed 18 SSCg models having a variable number of classes. SSCg species classes that contained fewer than 161 images were added to an "other species" class. The exception to this rule occurred when a genus contained exactly two species (e.g., *Echinorhinus* and *Negaprion*). In this case, regardless of training data quantity, both species were trained with their respective labels, and the "other species" class was not incorporated. Regardless of species-specific dataset size, if their parent genus did not meet the threshold, a SSCg local model was not trained. We trained 18 SSCg models with 19,243 images. The SC is capable of classifying 47 species.

We optimized the models with the categorical cross-entropy function to learn shark features that are specific to genus and species classes [95]. To prevent redundant feature training

and incorporate fewer parameters, we used Densely Connected Convolutional Network (201 layers) (DenseNet201) as our pre-trained network for multiclass classification [70]. We used image augmentations and passed our training dataset through 20 convolutional-pooling layers to generate feature maps at each layer, creating weights. We adjusted the SC's learning rate to $9 \times 10^{-3}$ with the Adaptive Gradient Algorithm (Adagrad) optimizer [41]. We activated layers with the sigmoid function [114], which in the case of the SC facilitated higher test accuracy than ReLU activation units. We normalized classes with the softmax activation function [21]. We tuned dropout to 15% and trained the models over 15 epochs while incorporating early stopping if validation loss increased for five consecutive epochs [144].

### 2.2.4 Shark Detector Performance

To demonstrate the potential of this approach, we applied the SD to three common cases of data generation methods involving sharks and measured performance. In the first method, we used the SL to locate shark subjects in the sharkPulse dataset. We used a detection threshold of 0.9. Before locating shark subjects, we resized images to $512 \times 512$. We calculated the proportion of shark images that were located by the SL.

Second, we evaluated a data pipeline we developed for sharkPulse, where global sighting records of sharks are generated from the SN IG [52]. We extracted images from IG, identified shark images (Table 2.2), and classified them (Table 2.3). In this case, we excluded the SL when processing IG images because it was more computationally expensive and did not significantly impact classification accuracy.

In the third method, we post-processed two BRUV videos and five YT videos with the goal of locating, identifying, and classifying all sharks present (Figure 2.2, Table 2.4). Videos were chosen to represent varying habitat types, conventional ecological surveys, and data-

mining methods focused on sharks. We processed all videos sequentially to evaluate total processing time in addition to individual processing time. All videos were recorded at 30 frames per second. We extracted the first frame per second and resized frames to $512 \times 512$ to reduce memory consumption and decrease processing time. The extracted frames were then screened with the SL. The SL threshold was increased to 0.99 to minimize the False Positive (FP) rate. We tested the SL to identify all sharks per frame. To test the SL's specificity on a video that did not contain any shark subjects, we processed a YT video that exclusively depicted typical coral reef habitats and numerous fish species, but no sharks (see YT Video 2 in Table 2.4). We annotated shark-located frames with the timestamp in the video from which they were extracted. This was done so we could check the video at the exact time a shark was located. We sorted shark-located frames with the SI (threshold 0.5). Finally, the identified shark images were classified with the SC. We calculated the SC's recall for its top guess and its top three guesses.

## 2.3 Results

### 2.3.1 Boosting Training Data

The SD components performed well individually and as a stepwise process. Overall, the SL performed at 89% accuracy, the SI at 91% accuracy, and the SC at 69% accuracy (the SCv5 performed at 80%). The SL located 90% of shark images from the sharkPulse data archive ($n$ = 24,546 shark images checked by manual review) and generated novel training data by extracting only shark features. By locating one or multiple subjects in shark images (Figure 2.2), the SL cropped 28,799 additional images from videos to inflate the original training dataset. We taxonomically labeled 14,888 cropped images to the genus level and

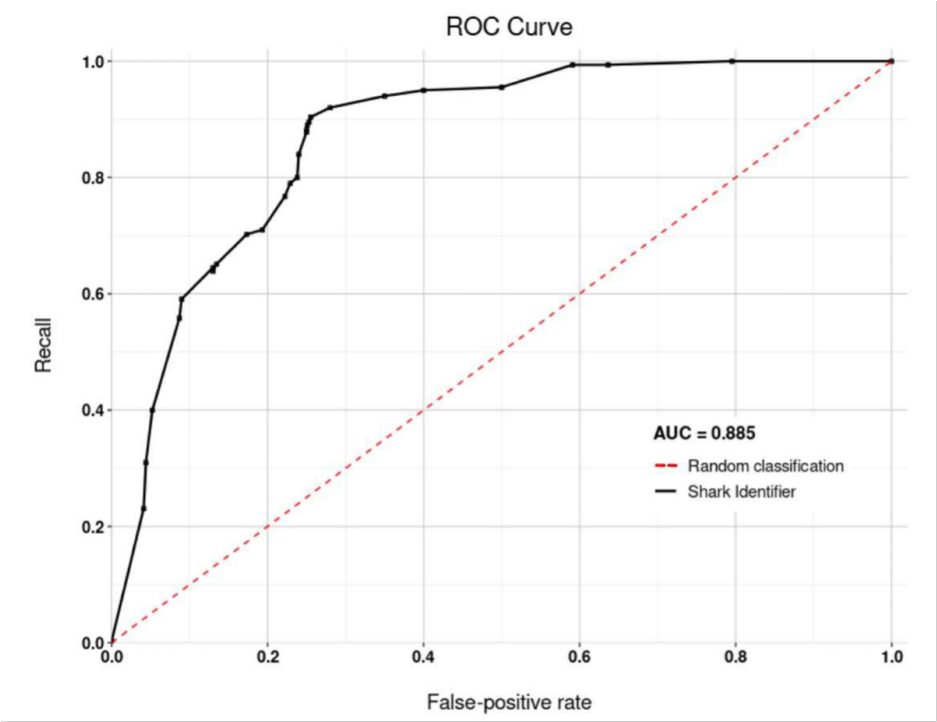


Figure 2.3: Receiver Operating Characteristic (ROC) curve of the SI binary classification scheme. The Area Under the Curve (AUC) conveys a probability measure of how likely the model is to separate between positive and negative classes. The red, dotted diagonal line indicates a no-skill classification model that discriminates randomly.

9,979 images to the species level for inflating the GSC and SSCg respectively. All subsequent SD models dramatically increased their accuracy as a result of ingesting this training data. We observed an average 3.5% increase in test accuracy of all models that used training datasets inflated by the SL.

### 2.3.2 Training and Performance

By examining the training data threshold distribution of GSC and SSCg classes, it was revealed that $433 \pm 47$ images are needed to achieve $> 50\%$ recall (above random classification) among genus classes and $161 \pm 41$ images are needed to achieve the same recall among SSCg classes (see Figures 2.4c–d). However, variability across genera and species is high in Figures 2.4a–b, and the relationship between the two variables depends on morphological distinctiveness as well as the level of training data balance. For instance, morphologically distinct species such as *Rhincodon typus* (Euclidean distance to centroid 19.0) and *Orectolobus* spp. (distance 25.6) required significantly fewer training images than species with common physical attributes such as *Carcharhinus* spp. (distance 8.3) and *Prionace glauca* (distance 9.0). We calculated the Pielou diversity index of the GSC to be 0.94 (scale 0–1), meaning genus training datasets were overall well balanced. The average Pielou diversity index of SSCg models was 0.77. Lastly, we compared the classification accuracy of iNat's classifier *Seek* with the SD on 400 random shark images sourced from IG. There were 13 species to classify. *Seek* performed at 62% top classification $F_1$ Score ($F_1$) while the SD performed at 73% top $F_1$.

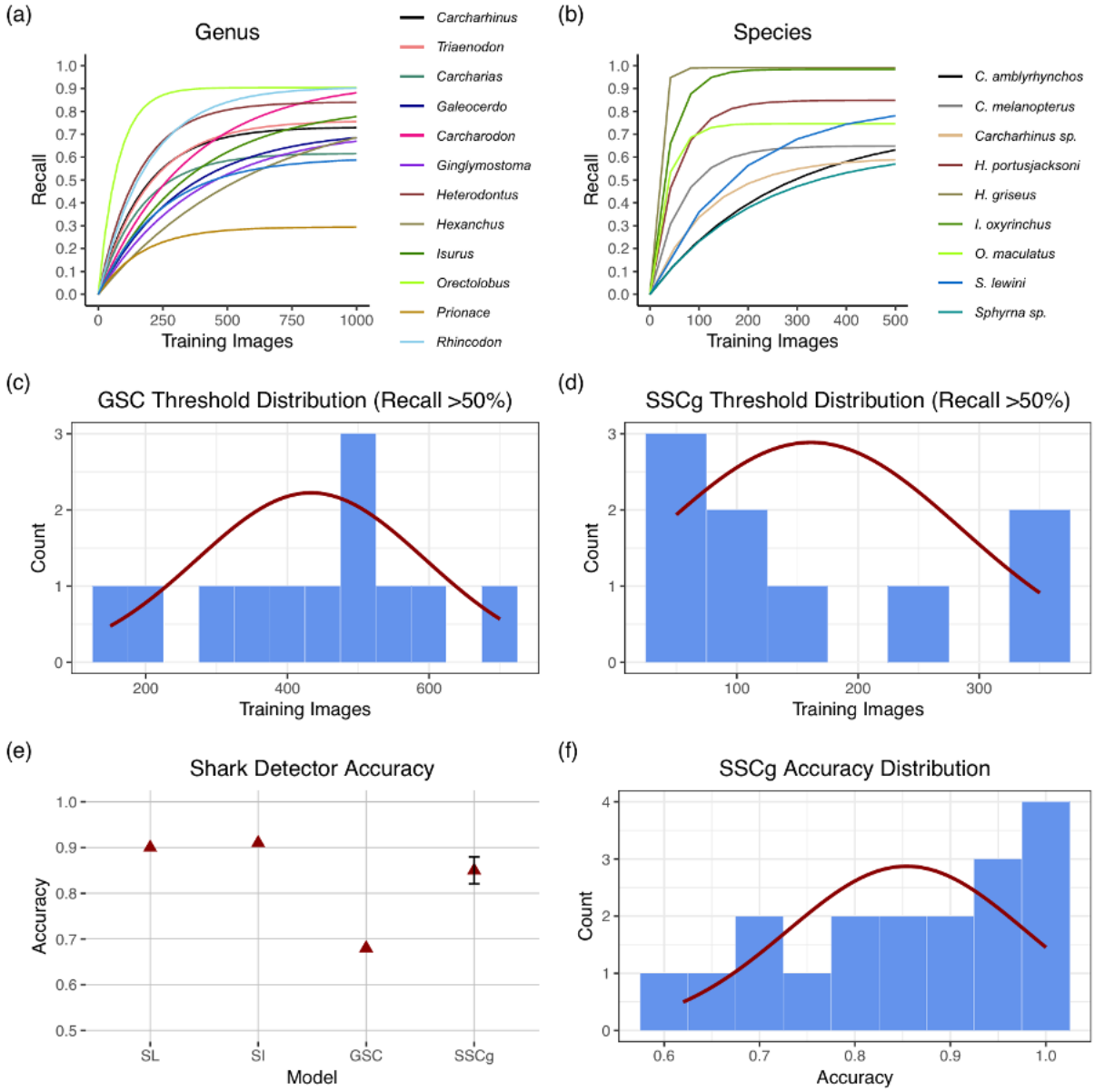


Figure 2.4: Measured performance of SD components. (a) GSC accuracy for 13 genus classes as a result of training dataset size fit with a two-parameter asymptotic model. The asymptotic curves represent the maximum recall a class can achieve within the model. (b) SSCg accuracy of seven species classes and two classes that contain a data-poor *Carcharhinus* sp. and *Sphyrna* sp. (c) Distribution of dataset size threshold for 12 GSC classes (*Prionace* was excluded due to recall never reaching 50%). Curves represent the density of a normal distribution. (d) Distribution of dataset size threshold for nine SSCg classes whose parent genera contain more than two species. (e) Performance of all SD components with a standard error interval for SSCg models. (f) Accuracy distribution of all 18 SSCg models.

### 2.3.3 Instagram

By data-mining images from IG, we created 14 datasets (Appendix B). The SI removed non-shark images and retained shark images with 91% overall accuracy (Table 2.2, Appendix B.1). About 5% of actual shark images were not related to the hashtag they were scraped from (i.e., they were other species) (Appendix B.2). The area under the correlation between recall and False Positive Rate (FPR) of the SI represented a successful classification probability of 0.885 (Figure 2.3). The SI displayed lower recall when sorting video frames that had not yet been cropped for shark subjects but performed well when sorting heterogeneous data-mined images. The SI displayed a low FPR and False Negative Rate (FNR). However, we noticed images like those in Figure 2.5b and 2.5c represented commonly misclassified images. These misclassifications occurred 9% of the time.

| Hashtag | Test | TP | FP | TN | FN | Recall | Precision | Specificity | FPR | FNR | $F_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| #tigershark | 1590 | 299 | 19 | 1261 | 11 | 0.96 | 0.94 | 0.99 | 0.01 | 0.04 | 0.95 |
| #blueshark | 1269 | 343 | 22 | 889 | 15 | 0.96 | 0.94 | 0.98 | 0.02 | 0.04 | 0.95 |
| #whaleshark | 1144 | 309 | 21 | 800 | 14 | 0.96 | 0.94 | 0.97 | 0.03 | 0.04 | 0.95 |
| #makoshark | 988 | 228 | 15 | 730 | 15 | 0.94 | 0.94 | 0.98 | 0.02 | 0.06 | 0.94 |
| #scallopedhammerhead | 871 | 149 | 18 | 701 | 3 | 0.98 | 0.89 | 0.97 | 0.03 | 0.02 | 0.93 |
| #sandtigershark | 1011 | 220 | 21 | 753 | 17 | 0.93 | 0.91 | 0.97 | 0.03 | 0.07 | 0.92 |
| #nurseshark | 1370 | 180 | 23 | 1156 | 11 | 0.94 | 0.89 | 0.98 | 0.02 | 0.06 | 0.91 |
| #greatwhites | 849 | 251 | 24 | 550 | 24 | 0.91 | 0.91 | 0.96 | 0.04 | 0.09 | 0.91 |
| #blacktipshark | 955 | 209 | 40 | 700 | 6 | 0.97 | 0.84 | 0.95 | 0.05 | 0.03 | 0.90 |
| #portjacksonshark | 1012 | 191 | 28 | 971 | 13 | 0.94 | 0.87 | 0.97 | 0.03 | 0.06 | 0.90 |
| #sixgillshark | 287 | 107 | 19 | 150 | 11 | 0.91 | 0.85 | 0.89 | 0.11 | 0.09 | 0.88 |
| #spottedwobbegong | 180 | 97 | 18 | 55 | 10 | 0.91 | 0.84 | 0.75 | 0.25 | 0.09 | 0.87 |
| #whitetipreefshark | 890 | 250 | 43 | 561 | 36 | 0.87 | 0.85 | 0.93 | 0.07 | 0.13 | 0.86 |
| #greyreefshark | 901 | 203 | 55 | 600 | 43 | 0.83 | 0.79 | 0.92 | 0.08 | 0.17 | 0.81 |
| total | 13317 | 3036 | 366 | 9877 | 229 | 0.93 | 0.89 | 0.96 | 0.04 | 0.07 | 0.91 |

Table 2.2: Hashtags relevant to specific shark species that were data-mined from IG. The result was heterogeneous datasets of images. We measured the SI's sorting accuracy and error rate at a confidence threshold of 0.5.

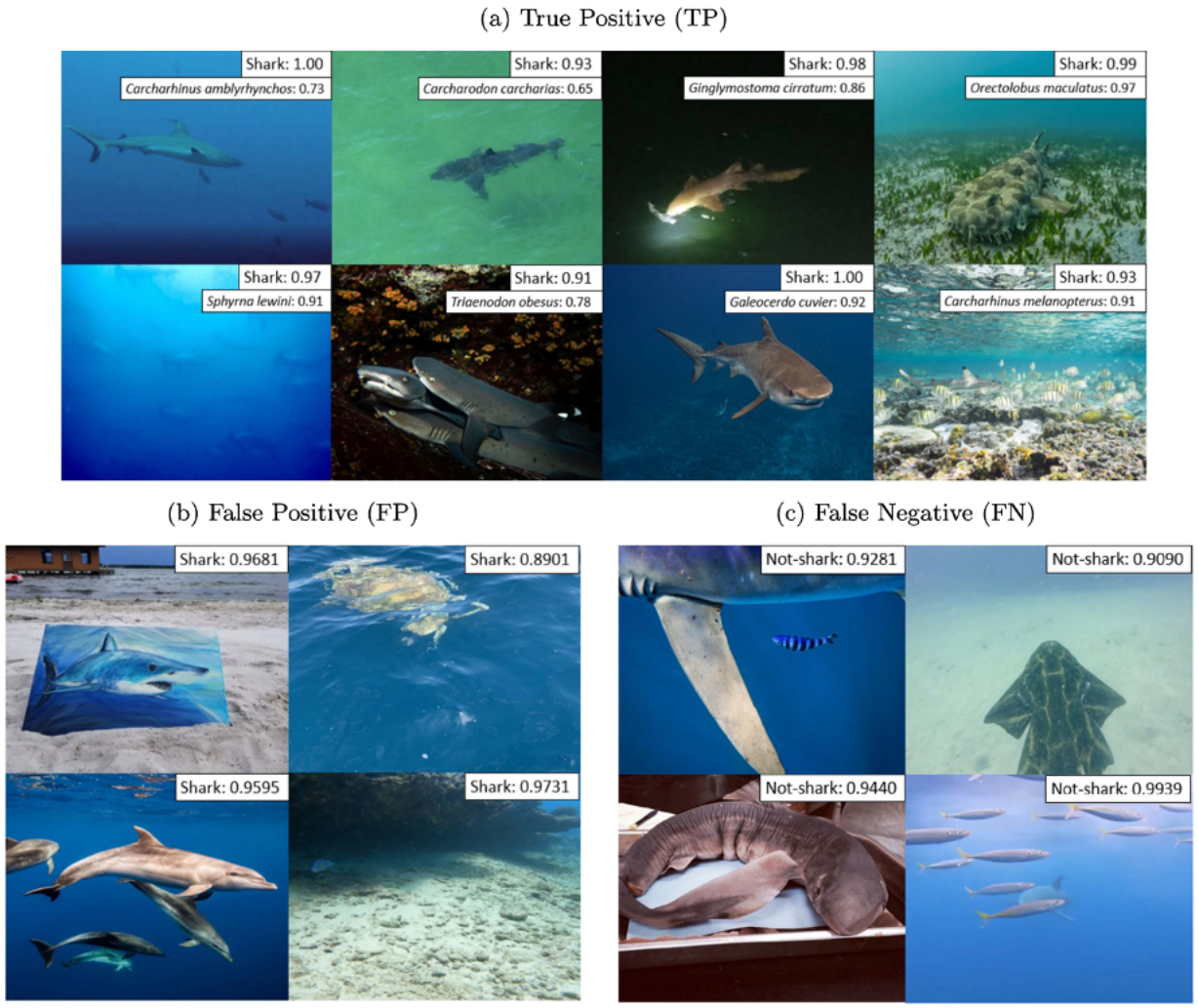


Figure 2.5: Images identified by the SI and subsequent classification by the SC. (a) The SI and SC correctly identify a diverse collection of shark images by classifying underwater photographs, images with foreground and background noise, images with hardly discernible shark features, and eight different species. (b) Common subjects that were misclassified by the SI such as cetaceans (and other marine and terrestrial animals), empty foregrounds, inscrutable objects, and artificial models. (c) The SI misses shark presence due to partially concealed features.

| Species | Scientific Name | Training images | Test images | Recall | Top-3 Recall |
|---|---|---|---|---|---|
| Whale shark | *Rhincodon typus* | 1602 | 309 | 0.95 | 0.99 |
| Port jackson shark | *Heterodontus portusjacksoni* | 1172 | 191 | 0.87 | 0.98 |
| White shark | *Carcharodon carcharias* | 2290 | 251 | 0.87 | 0.90 |
| Whitetip reef shark | *Triaenodon obesus* | 1786 | 250 | 0.79 | 0.92 |
| Blacktip reef shark | *Carcharhinus melanopterus* | 829 | 209 | 0.77 | 0.91 |
| Shortfin mako | *Isurus oxyrinchus* | 1360 | 228 | 0.76 | 0.95 |
| Spotted wobbegong | *Orectolobus maculatus* | 1019 | 97 | 0.74 | 0.99 |
| Nurse shark | *Ginglymostoma cirratum* | 821 | 180 | 0.70 | 0.90 |
| Tiger shark | *Galeocerdo cuvier* | 1117 | 299 | 0.68 | 0.91 |
| Grey reef shark | *Carcharhinus amblyrhynchos* | 550 | 203 | 0.68 | 0.88 |
| Bluntnose six-gill shark | *Hexanchus griseus* | 792 | 107 | 0.68 | 0.71 |
| Sand tiger shark | *Carcharias taurus* | 2405 | 220 | 0.67 | 0.89 |
| Scalloped hammerhead | *Sphyrna lewini* | 274 | 149 | 0.60 | 0.84 |
| Other species | – | 1086 | 88 | 0.50 | 0.71 |
| Blue shark | *Prionace glauca* | 990 | 343 | 0.29 | 0.76 |
| Total | – | 18093 | 3124 | 0.70 | 0.90 |

Table 2.3: SC classification of data-mined images from IG. Recall was measured for the SC's top species prediction as well as the top three predictions.

### 2.3.4 BRUV Surveys and Online Videos

We classified eight shark species from seven videos. We processed two BRUVs recordings and five YT videos that made up 136 minutes of total video footage, which contained eight species of sharks. We spent 6.2 hours manually validating all of the extracted frames ($n$ = 8,185 frames). It took the SD 2.6 hours to process all videos in succession. The SL located 89% of available shark frames ($n$ = 2,277 frames) and the SI filtered out false positive images with 94% specificity (Table 2.4). The SC classified all species with an average top recall of 69% and top-3 recall of 76%. The SL showed 93% specificity and a false-positive rate of 7% when processing YT Video 2, which did not contain sharks.

| Metric | Sicilian Channel | Palau Archipelago | YT 1 | YT 2 | YT 3 | YT 4 | YT 5 |
|---|---|---|---|---|---|---|---|
| Video length (min) | 35.4 | 17.7 | 49.2 | 10.5 | 4.3 | 14 | 5.6 |
| Processing time (min) | 37.1 | 18.2 | 55 | 13.5 | 6.2 | 17.2 | 8.0 |
| Frames extracted | 2121 | 1055 | 2951 | 630 | 255 | 841 | 333 |
| # of shark images | 812 | 152 | 855 | 0 | 120 | 256 | 82 |
| # of non-shark images | 1309 | 903 | 2096 | 630 | 135 | 585 | 251 |
| SL Recall | 0.90 | 0.88 | 0.89 | 0.80 | 0.90 | 0.90 | 0.91 |
| SL Precision | 0.91 | 0.84 | 0.90 | 0.93 | 0.93 | 0.86 | 0.87 |
| SL Specificity | 0.92 | 0.85 | 0.87 | 0.93 | 0.89 | 0.84 | 0.84 |
| SI Specificity | 0.97 | 0.94 | 0.90 | 0.95 | 0.96 | 0.96 | 0.96 |
| SC Recall | 0.62 | 0.79 | 0.69 | 0.78 | 0.82 | 0.82 | 0.73 |
| SC Top-3 Recall | 0.70 | 0.86 | 0.76 | 0.84 | 0.88 | 0.88 | 0.81 |

Table 2.4: Performance metrics of SD components to locate, identify, and classify sharks from two BRUVs and five YTs videos that collectively depict eight species of sharks. SL threshold 0.99, SI threshold 0.5.

Similarly, we assessed the recall of species classes within SSCg models (see Figure 2.4b). Both the *Hexanchus* and *Isurus* models contained two unbalanced classes (see Table 5 for training datasets). So, we anticipated that the recall of these models' dominant classes would peak even if they were trained with fewer images. The models *Carcharhinus*, *Heterodontus*, *Orectolobus*, and *Sphyrna* contained mostly balanced training datasets with four or more classes. Interestingly, port jackson shark (*Heterodontus portusjacksoni*), spotted wobbegong (*Orectolobus maculatus*), and blacktip reef shark (*Carcharhinus melanopterus*) classes reached their maximum recall while being trained with fewer than 200 training im-

ages. Grey reef shark (*Carcharhinus amblyrhynchos*), other *Carcharhinus* sp., scalloped hammerhead (*Sphyrna lewini*), and other *Sphyrna* sp. classes did not meet their maximum recall with fewer than 500 training images.

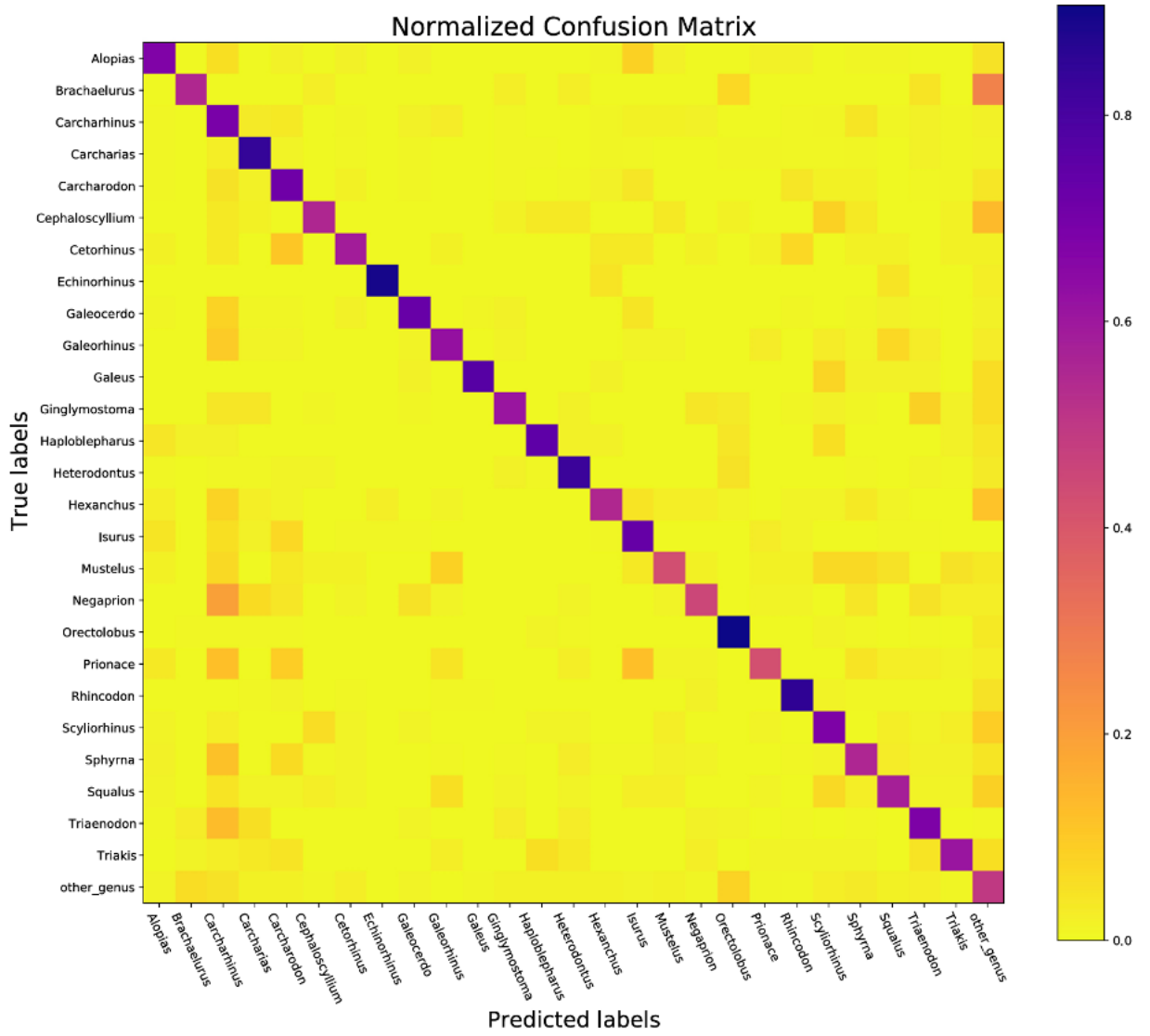


Figure 2.6: GSC normalized confusion matrix of 26 shark genera classes. A 27th class "other genus" represents 48 data-deficient genera.

| Species | Images | Accuracy | Species | Images | Accuracy | Species | Images | Accuracy |
|---|---|---|---|---|---|---|---|---|
| ***Alopias*** | 1185 | 0.65 | ***Galeorhinus*** | 791 | 0.53 | ***Orectolobus*** | 2021 | 0.92 |
| *A. vulpinus* | 353 | 0.81 | *G. galeus* | 791 | 1.00 | *O. maculatus* | 1019 | 0.82 |
| *Alopias spp.* | 174 | 0.72 | ***Galeus*** | 575 | 0.72 | *O. halei* | 542 | 0.62 |
| ***Brachaelurus*** | 479 | 0.65 | *G. melastomus* | 376 | 1.00 | *O. ornatus* | 281 | 0.34 |
| *B. waddi* | 299 | 0.96 | ***Ginglymostoma*** | 945 | 0.61 | *Orectolobus spp.* | 97 | 0.60 |
| *B. colcloughi* | 162 | 1.00 | *G. cirratum* | 821 | 1.00 | ***Prionace*** | 990 | 0.42 |
| ***Carcharhinus*** | 4963 | 0.71 | *G. unami* | 124 | 0.92 | *P. glauca* | 990 | 1.00 |
| *C. melanopterus* | 829 | 0.70 | ***Haploblepharus*** | 680 | 0.61 | ***Rhincodon*** | 1602 | 0.88 |
| *C. amblyrhynchos* | 550 | 0.67 | *H. fuscus* | 271 | 0.96 | *R. typus* | 1602 | 1.00 |
| *C. limbatus* | 488 | 0.41 | *H. edwardsii* | 215 | 0.23 | ***Scyliorhinus*** | 964 | 0.63 |
| *C. leucas* | 402 | 0.66 | *Haploblepharus spp.* | 194 | 0.75 | *S. canicula* | 378 | 0.97 |
| *C. obscurus* | 259 | 0.85 | ***Heterodontus*** | 2180 | 0.82 | *Scyliorhinus spp.* | 94 | 0.50 |
| *C. perezi* | 245 | 0.24 | *H. portusjacksoni* | 1172 | 0.94 | ***Sphyrna*** | 1591 | 0.54 |
| *C. plumbeus* | 212 | 0.45 | *H. galeatus* | 343 | 0.89 | *S. tiburo* | 377 | 0.79 |
| *Carcharhinus spp.* | 815 | 0.74 | *H. francisci* | 337 | 0.74 | *S. lewini* | 274 | 0.82 |
| ***Carcharias*** | 2405 | 0.84 | *H. japonicus* | 306 | 1.00 | *S. mokarran* | 165 | 0.47 |
| *C. taurus* | 2405 | 1.00 | *Heterodontus spp.* | 22 | 0.84 | *Sphyrna spp.* | 140 | 0.79 |
| ***Carcharodon*** | 2290 | 0.72 | ***Hexanchus*** | 971 | 0.67 | ***Squalus*** | 1044 | 0.52 |
| *C. carcharias* | 2290 | 1.00 | *H. griseus* | 792 | 1.00 | *S. acanthias* | 182 | 1.00 |
| ***Cephaloscyllium*** | 663 | 0.56 | *Hexanchus spp.* | 8 | 0.00 | *Squalus spp.* | 130 | 0.77 |
| *C. isabellum* | 323 | 1.00 | ***Isurus*** | 1636 | 0.72 | ***Triaenodon*** | 1786 | 0.69 |
| *C. laticeps* | 264 | 1.00 | *I. oxyrinchus* | 1360 | 0.99 | *T. obesus* | 1786 | 1.00 |
| *Cephaloscyllium spp.* | 76 | 0.80 | *I. paucus* | 62 | 0.25 | ***Triakis*** | 1060 | 0.59 |
| ***Cetorhinus*** | 642 | 0.57 | ***Mustelus*** | 677 | 0.43 | *T. semifasciata* | 673 | 1.00 |
| *C. maximus* | 642 | 1.00 | *M. canis* | 187 | 0.68 | *T. megalopterus* | 213 | 0.86 |
| ***Echinorhinus*** | 516 | 0.87 | *Mustelus spp.* | 335 | 0.88 | *T. scyllium* | 161 | 1.00 |
| *E. cookei* | 452 | 1.00 | ***Negaprion*** | 910 | 0.38 | *Triakis spp.* | 12 | 0.00 |
| *E. brucus* | 6 | 0.00 | *N. brevirostris* | 171 | 1.00 | | | |
| ***Galeocerdo*** | 1117 | 0.72 | *N. acutidens* | 69 | 0.00 | | | |
| *G. cuvier* | 1117 | 1.00 | | | | | | |

Table 2.5: List of species, and number of training images, for which we could infer a taxonomic identification at the genus level (with the GSC model) and species level (with the SSCg models).

## 2.4 Discussion

Historically hampered by problems of data paucity, shark research is transitioning toward a time with ubiquitous big data. Embracing this movement requires being able to capture and structure the increasing amount of information available online and generated by modern scientific monitoring. In this context, we developed a modular software package targeted at identifying and classifying shark images from unstructured and unlabeled media. In this package, location and identification models were able to detect sharks with 90% and 91% recall, respectively. Further, a pseudo-hierarchical classification structure classified 26 genera and 47 shark species, at 69% and an average of 85% recall, respectively. Trained on the largest and most diverse shark image dataset compiled so far, this software facilitates rapid data collection on sharks and generation of biologically relevant data, including boosting information for data-poor species.

The full potential of this approach lies in achieving a completely automated data analysis pipeline. We have shown that surveys and online archives can be automatically processed for shark classification, although with human review needed. While the SD is moving toward complete automation, there is still room for improvement. Our shark detector is currently the most efficient software for locating, identifying, and classifying sharks from unlabeled media. The SD top species predictions were 11% more accurate than iNat's *Seek* (currently the best general-purpose biodiversity classifier available) [69]. Yet model accuracy can still be improved, especially for the GSC model.

The GSC acts as the parent node for multiclass classification among the SD components and is, therefore, most challenged in the pipeline (see Figure 2.4e). The GSC typically displayed lower classification error with more training samples. However, misclassification also depended on physical distinctiveness and training data balance and content. For ex-

ample, carpet sharks (*Orectolobus* spp.) are easily identified because they are physically unique. When comparing morphological Euclidean distances with all species represented, *Orectolobus* species exhibited one of the highest distances (25.6), meaning they are among the most physically dissimilar taxa. We also noticed the content of *Orectolobus* species' image and video archives were homogeneous because they are strictly bottom-dwelling sharks and are almost exclusively observed in benthic habitats. Therefore, the class achieved a high recall (92%) with $< 1000$ training images (see Figure 2.4a). Conversely, blue sharks (*Prionace glauca*) exhibit similar morphometric measurements to the centroid of the training dataset (Euclidean distance 9.0). They are frequently observed in various marine habitats by photographers, divers, and recreational and commercial fishers [26], resulting in heterogeneous image and video archives. As we continue to capture this heterogeneity and physical distinctiveness by gathering more images, we expect classification accuracy to increase. But currently, *P. glauca* experiences low recall (29%) with $< 1,000$ training images.

During training of the GSC, 15% of test images ($n = 539$ images) were mistaken for the *Carcharhinus* genus (trained with 4,963 images and representing 24 species) and the "other genus" class (trained with 2,593 images and representing 48 genera and 172 species). Because the *Carcharhinus* and "other genus" classes describe 196 species, their training datasets are heterogeneous and variable in morphology, imbalanced relative to other smaller genera, and represent 21% of the entire training dataset. While the GSC training dataset was shown to be well-balanced with Pielou's diversity index of 0.94, we can minimize confusion and improve overall classification accuracy by continuing to balance data-poor genera. Furthermore, boosting genera that do not reach the training threshold would remove them from the "other genus" label, reduce confusion with the label, and increase the SC's taxonomic range. The SC will gain a new classifiable genus capable of achieving $> 50\%$ recall. However, morphological diversity will still affect the GSC's overall training accuracy.

SSCg models are composed of child nodes that utilize previous taxonomic information from the GSC. As expected, average SSCg classification accuracy (85% with 3.5% standard error) was higher than GSC accuracy (Figure 2.4e–f). Nonetheless, even SSCg models were challenged by imbalanced datasets and class similarity. The *Hexanchus* and *Isurus* models each contain two classes, where the dominant class was trained with an average of 50 times more images than the non-dominant class. Recall was perfect for *Hexanchus griseus* and *Isurus oxyrinchus* (Figure 2.4b) because the model did not learn the misrepresented class. This affected our SSCg threshold distribution (see Figure 2.4d) by indicating fewer training images were needed to reach $> 50\%$ recall, without taking into account that the classes are imbalanced. Further, fitting asymptotic recall functions of different SSCg model classes was useful for gauging future data boosting efforts. For instance, we noticed a pattern where dominant classes attained their maximum recall ($< 500$ training images) while non-dominant classes did not. This suggests that maximum recall values are useful as benchmarks, but will change as species are boosted and classes are increasingly represented. To best increase overall SC top recall and species coverage (Table 2.3), we must grow the number of taxonomically labeled images while prioritizing data-poor species, balancing training datasets, and increasing image diversity.

SI/SL misclassifications are $< 10\%$ frequent. The Faster-R-CNN model allowed the SL to achieve high recall (89%), precision (88%), and specificity (93%) [123]. VGG16 allowed the SI to achieve an $F_1$ score of 91% [140]. Performance of the detection and classification models can be boosted by inflating training datasets.

The training and validation datasets are substantial considering the scarcity of visual information repositories for most shark species. sharkPulse contains the largest repository of shark images and provides a consistent influx of shark-specific media by combining several data collection approaches: data scraping from online archives, user submissions, and

synthetic image generation techniques. Our training data are high quality due to crowdsourcing validation of taxonomic and spatiotemporal information. This facilitates continuous data collection and classification accuracy and is slowly being adopted for conservation [52, 69, 103]. For example, iNat's *Seek* was trained on a massive database of crowdsourced images validated by the application's users. Effectively, iNat and *iSharkFin* grow with user submissions, which can improve the models' classification accuracy [9, 69]. We adopted this approach and combined it with automated data scraping and synthetic image generation techniques, making the SD a novel instrument for collecting visual media of sharks. While the SD excels at classification accuracy and taxonomic range compared to other methods, there are still objectives to strive for. *Seek* is available on smartphones and, as a result, can equip everyone with intelligent monitoring capabilities. Increasing citizen science interactions and validation effort with mobile applications would continue to improve the quality of data sourced from sharkPulse and the SD.

Utilizing unsupervised models for shark detection and species identification has multiple applications, including processing online videos, survey footage, and big data [139]. This allows us to expand possibilities for filling information gaps in shark populations, even beyond traditional fisheries monitoring techniques. IG is a massive data cloud that offers tremendous opportunities for generating biologically relevant data. However, it contains a daunting amount of irrelevant content that would be unrealistically filtered with manual validation [107]. Plus, even targeted shark images often lack taxonomic and spatiotemporal information that need to be inferred with postprocessing. When 91% of noisy data are removed, and the remaining content is taxonomically classified, validation suddenly becomes practical. Furthermore, filtering and classifying facilitate the development of geoparsing and time-stamping programs [107]. Preliminary investigations suggest that IG posts of sharks can be effectively transformed into occurrence records with these taxonomic and spatiotemporal

identifiers [52].

The largest limitation of the SD is classification accuracy for data-poor species. Boosting natural and synthetic image generation techniques will inflate the training datasets of these species considerably, and subsequently increase classification accuracy. We showed how data-mining (Table 2.2), object-detected cropping (Figure 2.2), and image augmentations are effective data generation approaches. SNs like IG offer an inexhaustible source of shark and non-shark images [52]. Furthermore, synthetic image generation can be significantly improved. We can extract cropped images of fish and paste them onto randomly selected backgrounds while incorporating transformations. This approach will effectively generate thousands of new images from a handful of genuine images [2]. As new shark images are ingested and validated, the SD will immediately use them, automatically funneling those images into the appropriate training datasets. The SD will be a rapidly evolving Artificial Intelligence (AI), automatically collecting and generating new shark images, training models, and growing smarter with each step.

Despite these opportunities, careful consideration of risks is essential before the SD can be widely adopted. One prominent concern is that automatic classification may unintentionally enable the targeting of vulnerable populations if precise locations or identifications are made public without safeguards. These issues may be exacerbated for species already subject to exploitation or harassment, and highlight the need for deliberate redaction of sensitive spatial information. Another risk lies in the potential misuse of the software by untrained users or in settings where data quality cannot be guaranteed. Low-resolution images, poor visibility, or intentionally misleading inputs can yield spurious predictions, and without appropriate safeguards these errors may propagate into scientific analyses or management decisions. Furthermore, machine learning models inevitably carry biases rooted in their training datasets. If uncorrected, these biases may lead to systematic over- or under-

estimation of certain species or regions, potentially skewing policy outcomes. Finally, the integration of the tool into regulatory or compliance frameworks must be done with caution. Automated predictions should complement, rather than replace, human verification, especially in high-stakes contexts such as enforcement of bycatch regulations. Addressing these challenges requires transparency about model limitations, active recalibration as new data are incorporated, and clear communication to both technical and non-technical users. By acknowledging and mitigating these risks, the SD can serve as a robust and responsible platform for advancing shark conservation while minimizing opportunities for misuse.

The SD has clear applications in policy and fisheries monitoring, particularly in contexts where traditional observer programs are limited by cost and human resources. Automated detection and classification can substantially reduce the burden of manual video review by rapidly identifying bycatch events in longline and trawl fisheries. This capability allows observer programs to generate standardized indices of shark encounters, such as Sightings per Unit Effort (SPUE), with greater efficiency and consistency than human annotation alone [23]. At the policy level, such indices provide valuable insight for management decisions, including the design of spatial closures, the timing of seasonal restrictions, and the evaluation of gear modifications intended to reduce shark mortality. The SD tool also has the potential to be deployed at landing sites or on vessels, where near real-time monitoring of deck footage could flag the capture of protected or prohibited species and trigger targeted compliance checks. Beyond regulatory applications, graphical interfaces of the SD could be adapted for fishers themselves. By providing automatic identifications, approximate size estimates from reference objects, and best-practice handling guidance, the tool can both enhance fisher decision-making and create an avenue for voluntary data sharing. Such applications tailored to fishers as a demographic could return value directly to the user, for example through digital catch logs or compliance support, while also generating occurrence records that can

feed into larger-scale monitoring frameworks. In this way, the SD not only complements existing observer programs but also expands the reach of monitoring to contexts where human expertise is limited [23].

The SD presented in this chapter represents the first version as the foundation of an evolving platform that has continued to expand in scope and functionality. In Chapter 2 and Chapter 3, we describe the $5^{th}$ version of the SD with key advances to the SL, SI, and SC components. First, we have since included the option to employ a BRUV-specific object detection model in the place of the SL or SI [164]. Second, we have increased the training dataset of the SI to nearly 500k images between shark and non-shark subjects, increasing its classification accuracy to 98%. Third, the species classification strategy has been extended from a single-step genus-to-species model to a conditional, hierarchical approach that progresses from order through family and genus to the species level. This structural refinement reduces misclassification and allows predictions to reflect the underlying taxonomic hierarchy. The taxonomic range has been substantially expanded, SDv5 now capable of classifying 80 species trained on a dataset of more than 200k images, representing a fourfold increase over the training material used in this chapter. Finally, the platform has grown beyond command-line models into accessible interfaces: a Graphical User Interface (GUI) and an accompanying R package have been developed to scale the tool's application across diverse user groups. Together, these improvements reflect the iterative development of the SD as both a research framework and a practical resource for conservation, ensuring that it remains adaptable and broadly useful as new data and technologies emerge.

Beyond its scientific and management applications, the SD also has value as an educational and outreach tool. Automated species recognition provides an accessible entry point for students and the public to engage with shark biology, computer vision, and conservation science as sharkPulse has demonstrated with interactive web application tools [52]. The

ability to process photographs or video clips through the software allows users to see first-hand how machine learning can transform heterogenous media into ecological information. In classroom or workshop settings, the SD can be used to demonstrate core principles of species identification, sampling bias, and basic machine learning accuracy assessments, while also exposing students to the possibilities and limitations of artificial intelligence in conservation. Further, the growing volume of annotated images and videos creates opportunities for developing three-dimensional reconstructions of sharks. These reconstructions could be used to estimate body size, swimming behavior, and generate realistic visualizations for training, research, or public engagement. By making these capabilities open-source and adaptable, the SD encourages both technical learning and broader awareness of the role that automated tools can play in modern wildlife monitoring.

This chapter establishes a generalizable, open, and scalable framework for turning variably-tagged visual media into ecological information. Methodologically, I combined object detection, hierarchical taxonomic classification, and diverse workflows to crowdsource heterogeneous data sources (BRUVs, citizen science SNs and platforms, and web videos). Scientifically, I provide automated media pipelines that deliver species-level occurrences at scales that were previously impractical. Practically, I discuss opportunities for enhancing monitoring programs (observer footage, port sampling, protected-area surveillance) and generating indices suitable for trend analysis, while recognizing uncertainty, bias, and data incompleteness. The approach is not shark-specific: it is a blueprint for any data-poor, visually observable taxon, and creates a foundation for integrating media-based indices with conventional assessments.

We developed and compiled object detection and classification models into a single, open-source package that applies transfer learning and deep CNNs to the challenge of shark recognition. To our knowledge, this represents the most reliable general-purpose identification

software for sharks, trained on the largest and most diverse image dataset assembled to date. By making the SD openly available (https://github.com/sharkPulse/Shark-Detector), we provide researchers, managers, and citizen scientists with a practical tool that can be adapted to new data sources and experimental needs. The primary aim of this package is to accelerate data collection for species and regions where information is scarce, while continuing to expand the taxonomic coverage and accuracy of the models as training data grow. In doing so, the SD demonstrates how automated identification can contribute to the broader big data revolution in ecology, filling persistent knowledge gaps and complementing traditional fisheries monitoring. Ultimately, identification without the bottleneck of manual validation has the potential to reshape the design, scope, and quality of studies on shark ecology, biology, and conservation, providing a foundation for data-driven conservation and management.

# Chapter 3

# Diversifying the Shark Detector

In Submission as J. Jenrette, A. Agustines, E. T. Spencer, R. Schallert, N. Arnoldi, D. Madigan, T. White, K. Koller, J. Berglund, D. Kinzer, B. Block, S. Khalid, and F. Ferretti. Diversifying visual detection and classification artificial intelligence for accessible, semi-automatic monitoring of sharks.

# Abstract

The capacity to monitor shark biodiversity is vital for conservation. Baited Remote Underwater Video Systems (BRUVs) are increasingly popular and cost-effective ways to boost this scientific effort. However, they remain hindered by extensive manual video annotation and a lack of accessible tools to streamline automated detection into broader ecological analyses. These challenges are exacerbated by complex backgrounds, variable visibility, and sparse encounters making species-specific indices difficult to label. To overcome these barriers, we tested an integrated suite of Artificial Intelligence (AI)-driven tools for automated shark detection and taxonomic classification. At the core is the Shark Detector (SD), which combines binary image classification, You Only Look Once (object detection algorithm) (YOLO)-based object detection, and a hierarchical species classifier trained on 264,712 images of 309 shark species. Paired with three open platforms: an Application Programming Interface (API), an R package (`sharkDetectoR`), and a lightweight desktop application (SharkByte), the updated SD is accessible across computational environments and user expertise. Tested on 46,332 holdout images and 13.9 hours of BRUVs footage from Palau and the Main Hawaiian Islands, these tools achieved 92% species-level accuracy as well as 94.7% detection accuracy, and reduced annotation time by up to 90%. Additionally, leveraging these open-source tools to rapidly generate 8,466 new species-tagged training images improved base species-specific recall by 7.7% and survey recall by 9.5%, illustrating the tangible benefits of targeted data augmentation. These findings underscore the value of combining scalable AI-driven classification with data-boosting efforts to advance shark biodiversity monitoring. By substantially reducing manual annotation demands while maintaining high detection and classification performance, this framework offers a practical solution to speed up BRUV post-processing. At the same time, it highlights automated methods as essential tools for conservation and management, especially to address data deficiencies in shark research.

## 3.1 Introduction

Sharks are integral to the balance of many marine ecosystems as top predators and mesopredators [40, 53, 125]. They are crucial ecological drivers and charismatic megafauna of cultural significance. However, global shark biodiversity is a growing concern, with 31.2% of all shark species considered threatened with extinction [45]. Yet, conservation and management efforts on sharks are hindered by knowledge gaps in species-specific distribution and abundance indices [82, 145].

To fill these knowledge gaps, non-invasive monitoring expeditions focus on collecting ecological data for characterizing shark biodiversity and habitat preferences while mitigating harmful environmental and biological impacts. These surveys typically employ Baited Remote Underwater Video Systems (BRUVs), Remotely Operated Underwater Vehicles (ROVs), and Environmental DNA (eDNA) analyses [33, 54, 80, 87, 139].

BRUVs are cost-effective platforms that allow expedited sampling across various marine habitat types and depths. They have many uses, including monitoring relative abundance and distribution, behavioral and growth traits, and species richness [27, 28, 60]. BRUVs surveys produce video footage that must be taxonomically annotated, requiring substantial manual labor [139, 167]. Species classification is especially challenging among cryptic, data-poor, and morphologically similar elasmobranchs [145]. Identifying sharks from underwater images can be challenging with diminished video resolution and light. Further, the animal's distinguishing morphological features may be hidden from view, especially if interaction with the recording apparatus is brief or far from view. Distinguishing features such as color patterns, spots, and distinctive size and fin shape reduce misidentifications. However, some families, such as Carcharhinidae and Squalidae ($n = 97$ species), share similar features that even with direct observation can lead to taxonomic misidentifications [22, 136]. Thus,

species-specific indices can be misinformed, prompting less effective management strategies and costly revaluations [130, 170].

Artificial Intelligence (AI) can be leveraged to increase taxonomic resolution while reducing manual labor [176]. Novel integration of AI into traditional visual surveys has paved the way for revolutionary improvements in data collection, processing efficiency, and ecological monitoring accuracy [169, 176]. BRUVs generate extensive ecological datasets, thus manually annotating them becomes time-consuming and prone to human error, creating bottlenecks in data processing [101, 152]. Recent advances in AI, particularly Convolutional Neural Networks (CNNs) trained for object detection and species classification, offer a powerful supplemental tool to mitigate these limitations by automatically filtering irrelevant content, detecting individual subjects, and classifying observations with high taxonomic resolution [91, 97, 142]. This automatic processing significantly reduces manual validation efforts, allowing for more rapid expert review and promoting expanded survey coverage [78, 164, 168].

However, the practical implementation of AI within visual surveys is notably influenced by several key BRUV characteristics. Factors such as video resolution, data rate, and lighting conditions directly affect image clarity [10, 60], impacting AI's ability to accurately detect and classify species [97]. Complex underwater environments featuring varied habitat structures, such as coral reefs or seagrass beds, introduce background noise and foreground obstructions, making automated detection more challenging. Furthermore, animal behavior, including rapid movement, distance from the camera, and variable positioning, can significantly impact detection accuracy, increasing false negative or false positive recognition rates [22]. Morphological similarities between species further complicate automated species classification, particularly under conditions of low visibility or compromised image quality common in marine settings [84, 162]. Despite these challenges, accessible, streamlined AI platforms continue to evolve, incorporating iterative model refinements that progressively

enhance scalability, efficiency, and accuracy of biodiversity assessments across diverse and complex survey conditions.

Training robust shark classification algorithms can be challenged by a scarcity of publicly accessible, high-quality, and ecologically tagged visual data [78]. Traditional fishery-dependent surveys provide limited visual archives of sharks, often associated with bycatch events [10]. Conversely, citizen science initiatives, independent research expeditions, and social network (SN)s collectively present vast, yet largely untapped reservoirs of shark sightings that can greatly enrich training datasets [52, 78]. Addressing this opportunity, sharkPulse is an integrative cyberinfrastructure designed to systematically absorb, filter, annotate, and warehouse global elasmobranch photo observations from diverse sources [52]. This expansive repository directly supports the Shark Detector (SD), a versatile AI-driven recognition and species classification tool, which is integrated into sharkPulse, powering its automated filtering and annotation capabilities demonstrated in Chapter 2, Section 2.2 [78].

Importantly, sharkPulse's infrastructure and its corresponding R package, `sharkPulseR`, and web portal (`http://sharkpulse.org`), are designed not only to benefit researchers through enhanced data processing and analysis but also to actively empower citizen scientists by providing simplified web tools for data submission and ecological annotation [52]. Currently, sharkPulse enables volunteers and researchers to contribute single-image observations and participate in annotating crowdsourced observations, enhancing the robustness and accuracy of the SD [78]. However, the infrastructure lacks systematic resources for contributors to upload entire video-processed datasets, which limits the potential for scaled automated data-ingestion, retraining, and refinement of the SD specifically for survey-based applications. By offering local video-processing tools with sharkPulse-submission functions, in addition to restructuring the SD to retrain submitted data and immediately provide the updated models through its own R package (with specific SD detection and classification

functions), users can directly enhance and utilize the tool's performance.

In this study, we aimed to streamline automatic monitoring of sharks in BRUVs footage by reducing the burden of manual validation and improving taxonomic classification. To achieve this, we refined and expanded the SD, integrating updated detection and classification modules into practical, accessible tools: the programmatic `sharkDetectoR` package and the lightweight desktop application SharkByte. These tools replicate the conventional steps of reviewing BRUV frames, detecting shark presence, identifying species, and logging annotations—but automate them to reduce effort and standardize outputs. Together, the tools enable efficient filtering of large video datasets, retention of shark observations, and direct submission of validated data to the sharkPulse cyberinfrastructure for iterative retraining and model improvement. We evaluated the tools on 46,332 sharkPulse images and 13.9 hours of BRUV footage from the Main Hawaiian Islands (MHIs) and Palau, and further demonstrated how augmenting regionally relevant species with additional video-sourced training data improved species-level performance (Appendix C). Here, we provide researchers and practitioners with field-ready, open-source AI-driven tools that make BRUV surveys more efficient and reproducible, while addressing the main limitations of automatically monitoring sharks at scale.

## 3.2 Methods

We developed an analytical pipeline for automatic shark detection and taxonomic classification, incorporating novel improvements from Jenrette et al. [73] and Chapter 2 to enhance accuracy and streamline post-processing of BRUV surveys. This pipeline was packaged in two complementary formats to serve different user needs. First, we created an R package (`sharkDetectoR`) that provides programmatic access to detection and classification func-

tions, suitable for researchers who require scalable integration into their workflows. Second, we developed a lightweight desktop application (SharkByte) that packages the same core functionalities into a graphical interface, enabling non-programmers or field practitioners, such as those working at sea without server access, to process videos locally with simple point-and-click commands. We evaluated these tools using a holdout sharkPulse image dataset and BRUV footage from two independent surveys (Figure 3.2), comparing classification accuracy, annotation speed, and sensitivity under variable environmental and video quality conditions. Finally, we demonstrated a data-boosting workflow in which users can specify region-specific shark species, process focal videos, and submit validated outputs to the sharkPulse server, triggering immediate retraining of the Shark Detector and iterative improvement of its performance.

### 3.2.1 Detecting and Classifying Sharks

To recognize sharks in videos, we used an object detection model, specifically with a You Only Look Once (object detection algorithm) (YOLO) architecture, which rapidly identifies focal subjects in a single pass [122]. The SD framework was originally designed with two independent binary models: an image classification CNN and a general YOLO model [78]. Both models are trained on a mix of above-water and underwater shark observations, integrated into sharkPulse and specifically optimized to filter large, heterogeneous datasets originating from SNs and online archives [52]. These models are accessible in the newly developed `sharkDetectoR` R package [75], offering diverse data processing advantages.

Given the relatively homogeneous nature of datasets derived from BRUV footage in this study we developed our analytical pipeline by integrating SharkTrack, a YOLO-based object detection model explicitly trained and optimized for underwater shark detection from BRUV

surveys [164]. SharkTrack's model weights, developed from 6,862 training images across 77 BRUV deployments at 25 global locations, achieve a detection accuracy of 84%. Hence, we used SharkTrack's detection capabilities of identifying shark subjects within video frames to automatically draw bounding boxes around each detected individual and crop out irrelevant background. This precise cropping enhanced downstream species-level classification accuracy ensured by the Shark Detector. We integrated SharkTrack at the top of the analytical pipeline, into both the `sharkDetectoR` package and the SharkByte local application, enabling processing and accurate taxonomic identification of sharks directly from BRUVs footage (Figure 3.1).

We further developed the SD with a hierarchical CNN for conditional taxonomic classification to identify the species of the cropped individual. The Shark Classifier (SC) submodule of the SD was built upon a MobileNetV2 backbone in PyTorch [129], which employs a structured framework where predictions are made at each of the four taxonomic levels: order, family, genus, and species. It begins by predicting the order level using a single-step (non-hierarchical) classification approach, after which each subsequent taxonomic level is predicted conditionally based on the classification of its parent (Figure 3.1). We implemented this conditional hierarchical approach to strategically improve classification accuracy by leveraging the inherent taxonomic structure of sharks, addressing common challenges posed by morphological similarity and limited visual distinctions among related taxa (Appendix D.1). We trained the SC on the Virginia Tech Advanced Research Computing (VT-ARC) cluster using two A100 GPUs, dropout for regularization, and weighted random sampling to address class imbalance [144]. We used the Adaptive Moment Estimation (Adam) algorithm to optimize training [89], with learning rate scheduling and light augmentation, training for up to 130 epochs to predict hierarchical taxonomy, with all model weights and metrics archived for deployment.

To train the SC, we leveraged the extensive sharkPulse dataset, which contains images of 309 shark species. To maintain a balanced training dataset, we incorporated 264,863 images describing 7 orders, 21 families, 38 genera, and 80 species with an imposed minimum-sample cutoff of 200 images for a species-level class to be trained, mitigating imbalanced classes [78]( Appendix D.2). We partitioned images with an 8:1:1 ratio into a train, validation, and test category respectively. These categories represent images needed to train the model and learn features, monitor the model's performance on images it has not yet seen to fine-tune overfitting, and evaluate the final trained model on new images to generate performance metrics. From the test dataset, we evaluated the accuracy at each taxonomic level independently, and the joint accuracy to classify from the order to the species level.

To support programmatic use, we built a Python Flask Application Programming Interface (API) [126] hosted on the sharkPulse server. The API provides callable services for core tasks, including object detection, binary shark vs. non-shark classification, and multi-species taxonomic classification (Figure 3.3). These services leverage a Graphics Processing Unit (GPU) for faster processing and enable users to run the detection and classification models independently or sequentially. The API directly communicates with the backend of the companion `sharkDetectoR` package, which was designed for researchers to easily interface with the models programmatically.

The `sharkDetectoR` R package [75] allows users to submit local media files for processing and to adjust the detection confidence levels for sensitivity (detection threshold). The user is given options to output bounding-box annotated frames and cropped images to visualize the explicit area of any given frame that is being identified as a shark and to remove background noise from potential training data. Output includes a structured annotations file including bounding box coordinates and species predictions with confidence scores (see Figure 3.4, Appendix C). Additionally, `sharkDetectoR` provides functions to retrieve the current SC model

performance metrics and total image contribution at each taxonomic level with `get_metrics`, and compile a distribution list of shark and ray species within a given geographic bounding box with `find_species`. This function takes an input geographic bounding box and produces a list of elasmobranch species that are most likely to be observed within the spatial boundary. We created this function to guide systematic data-boosting efforts, particularly for underwater surveys, providing users with an option to visualize what species are likely to be observed within the expected survey region by directly sourcing International Union for Conservation of Nature (IUCN) distribution assessments [71], mean distribution likelihoods from AquaMaps [86], and common depth ranges from FishBase [57].

### 3.2.2 The SharkByte Application

To further expand accessibility of SD functions for field deployment, we developed SharkByte, a Graphical User Interface (GUI) built with Python Qt version 5 (comprehensive Python bindings for the Qt framework) (PyQt5) [98] that integrates local detection and classification workflows. PyQt5 is a comprehensive set of Python bindings for the Qt v5 application framework, enabling Python to serve as an alternative to C++ for cross-platform application development [98].

SharkByte is tailored for efficiently processing high-resolution underwater video and designed to run on personal machines without requiring constant internet access or high-performance computing resources. This lightweight and intuitive application can be deployed on Windows and Mac operating systems and empowers field researchers to annotate and classify shark encounters directly from raw video files, bridging the gap between advanced machine learning and practical field-based monitoring of sharks.

To demonstrate SharkByte's real-world capacity to automatically process videos and re-

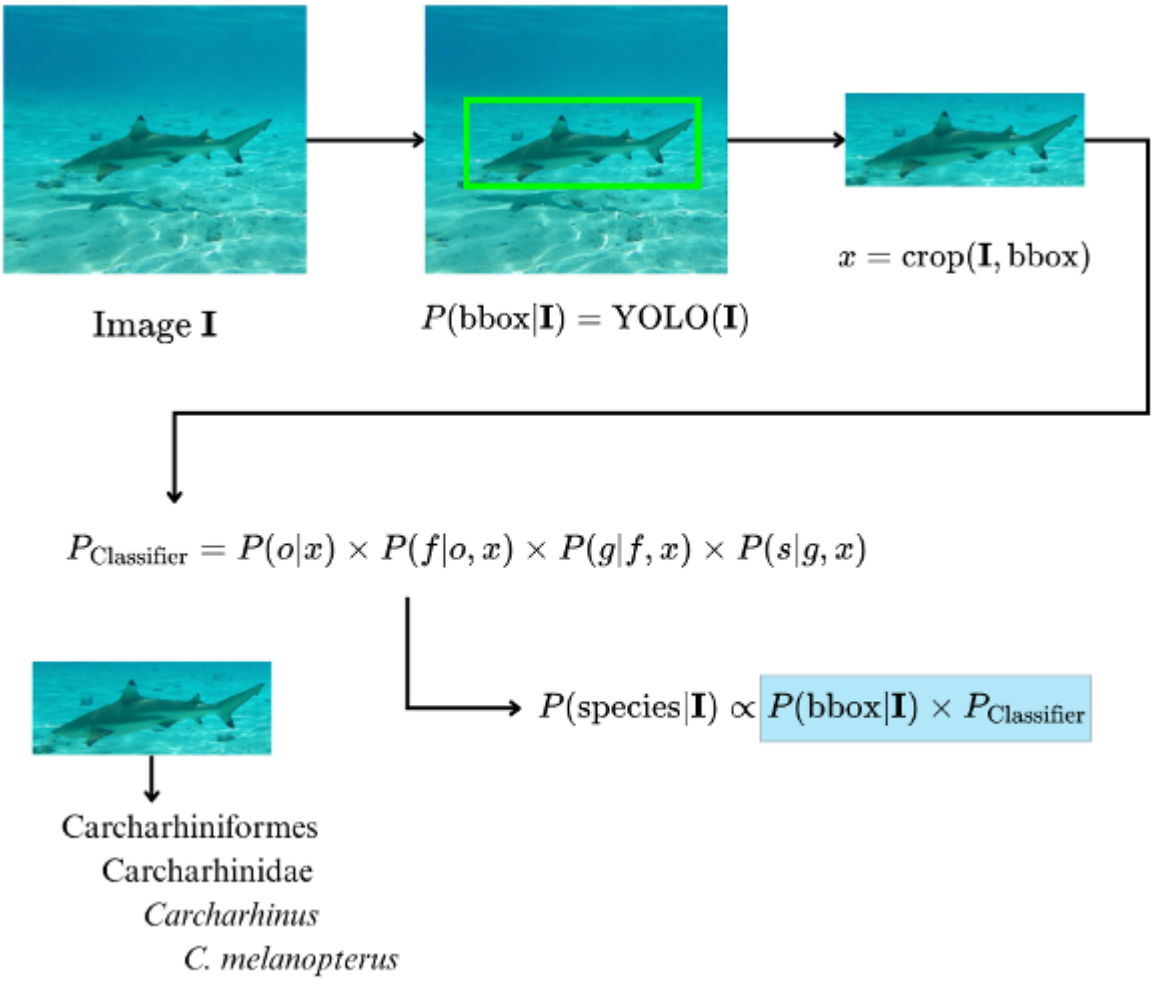


Figure 3.1: Workflow of the SDv5. An input image is first processed by a YOLO-based detector (SharkTrack) to locate sharks and generate bounding boxes. Cropped regions are then classified by a conditional CNN-based SC, which sequentially predicts order, family, genus, and species.

duce tedious manual annotation effort, we processed BRUVs footage from two independent surveys conducted in the MHIs and the Palauan Archipelago. Although differing in scope and design, both surveys aimed to visually characterize elasmobranch biodiversity, offering complementary settings to evaluate SharkByte's generalizable utility. The application enables users to input a single video, adjust detection thresholds, leverage local GPU hardware (if available), and control frame processing rate, including fractional frames per second, to balance speed with thoroughness (Figure 3.3). Output includes full video frames with bounding boxes, cropped images of detected individuals, an annotation file containing coordinates and species predictions with confidence scores, and an error log for debugging.

### 3.2.3 Integrating SharkByte into Biodiversity Surveys

We piloted the SharkByte software to post-process BRUVs footage from two demonstrative underwater surveys in the Palauan and Hawaiian archipelagos.

#### Palau Survey

In 2017, a Stanford team undertook an exploratory survey around the Palau Archipelago with the objective of assessing the shark community with multiple non-invasive approaches such as eDNA, BRUVs, and visual census scuba diving transects. The team deployed BRUVs from March 8th to March 13th. BRUVs platforms were deployed 10 m from the surface at, or in close proximity to, coral reef habitat (Figure 3.2A, Appendix E). The platforms consisted of an aluminum frame and two camera housings approximately one meter apart facing the same direction. The team recorded footage at 30 fps for 1–1.5 hours, accumulating 27 hours of footage across 11 deployments.

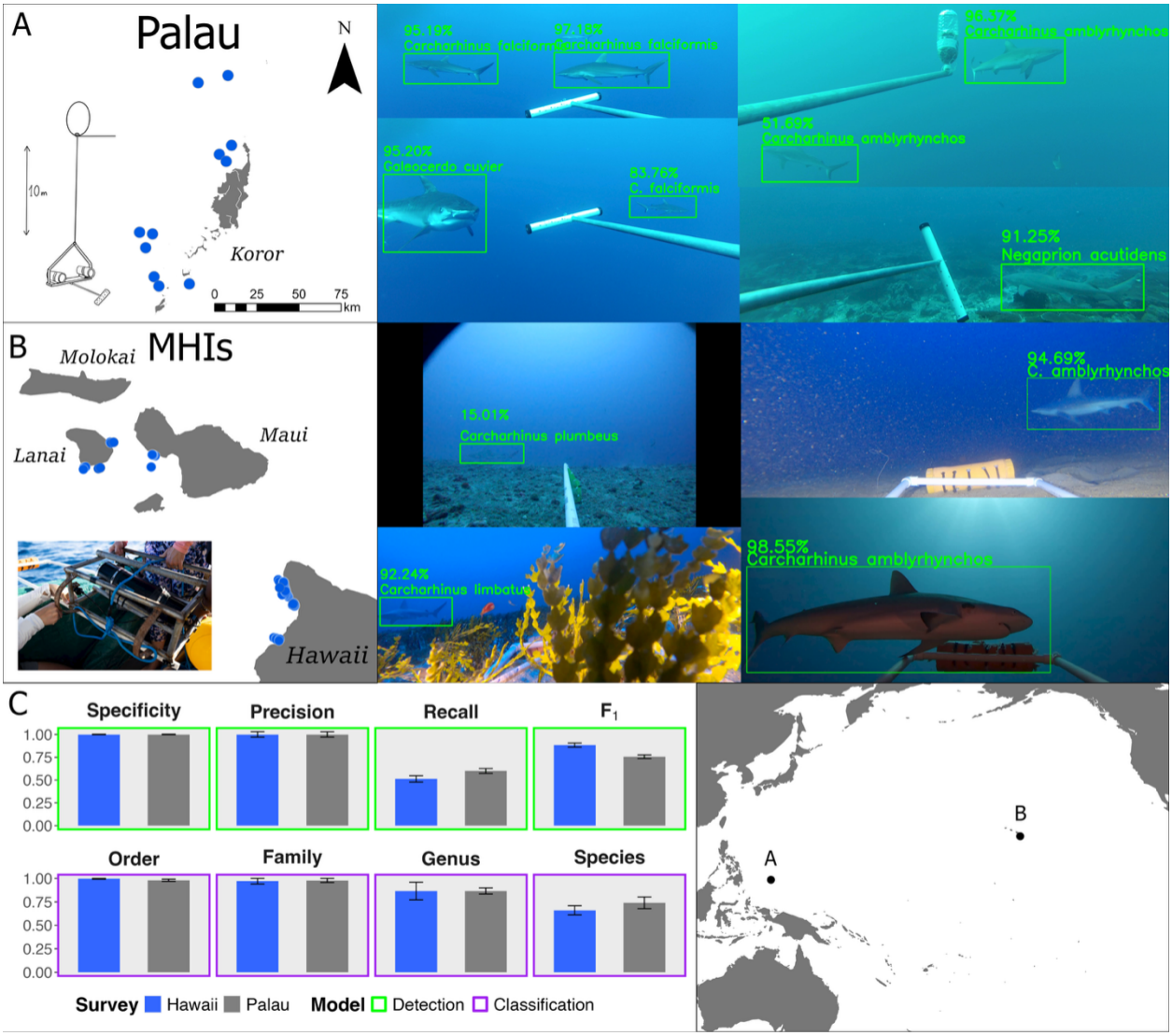


Figure 3.2: Automated shark detection and classification from BRUV surveys in two tropical regions. (A) Pelagic deployments around the Palauan Archipelago, with example frames showing detections of four species: gray reef (*Carcharhinus amblyrhynchos*), silky (*C. falciformis*), sicklefin lemon (*Negaprion acutidens*), and tiger shark (*Galeocerdo cuvier*). (B) Benthic deployments in the MHIs, with representative frames of gray reef, blacktip (*C. limbatus*), and sandbar shark (*C. plumbeus*): inset shows the benthic platform. (C) Performance metrics of detection: specificity, precision, recall, $F_1$ Score ($F_1$), and hierarchical classification (order to species) across survey regions. Inset map shows survey locations, illustrating the generalizability of the SD across ecological contexts.

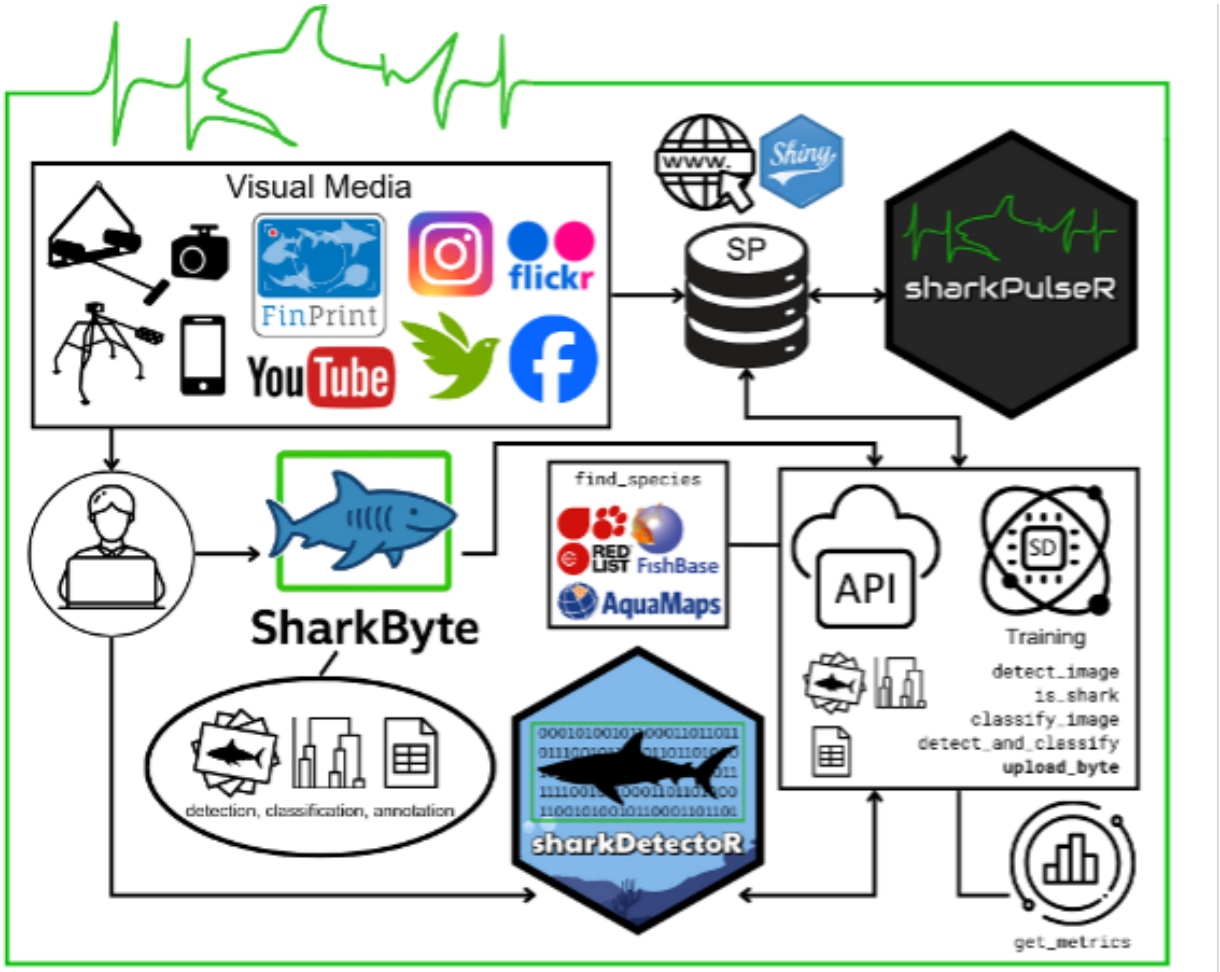


Figure 3.3: Illustration of the complete workflows connecting `sharkDetectoR`, SharkByte, and `sharkPulseR` within the sharkPulse cyberinfrastructure. Directional arrows indicate the flow of data or classification models. Connecting lines without arrows indicate a structural extension of a function such as `find_species`. The API and `sharkDetectoR` functions are indicated in normal font weight, while `upload_byte` indicates an API-specific function that only accepts compressed media submitted by SharkByte users. The sharkPulse relational database is shown as SP.

### Main Hawaiian Islands Survey

In 2022, the National Geographic Society partnered with the exploration vessel E/V *Nautilus* to launch the Mālama Manō Project (translation: Shark Protection Project), a multifaceted expedition with conservation and education-driven initiatives. An important function of the project was to assess the loss of elasmobranch biodiversity around the MHIs with non-invasive techniques such as eDNA and BRUVs analyses, integrating AI to post-process underwater footage. The expedition produced a total of 74 BRUVs deployments spanning 145 km around the MHIs, capturing 89 hours of video footage. The BRUVs surveys were conducted between 07:00 and 15:00 every day from September 17th to September 27th and October 2nd to October 8th, 2022. Surveys were completed across several locations around the islands of Lānaʻi, Maui, and Hawaii in shallow reef (2–40 m) and deep reef (41–250 m) habitats (Figure 3.2B, Appendix E). The BRUVs platform was designed for benthic deployments on the seafloor. The bait arm extended 1.5 m from the camera view and consisted of a PVC pipe with a mesh bag attached to one end where bait was loaded. To minimize ecological footprint, bait consisted of fish scraps sourced from local fish dealers. The platform settled on the seafloor and recorded a minimum of 60 minutes at 30 fps.

### Processing BRUVs

Both surveys recorded underwater video using GoPro Hero models 5 and above, which partition footage into multiple smaller files per deployment to mitigate data loss and manage file size. All output video files were processed independently using SharkByte. For consistency and comparability, we selected a subset of five deployments from each survey that had been manually reviewed by field teams and confirmed to contain shark observations. This approach ensured that the total number of deployments and the overall video duration were

matched across the two regions, allowing for standardized comparisons of annotation time and detection accuracy.

We processed a total of 13.8 hours of BRUVs footage—6.9 hours from each survey subset. The Palau subset contained 35 videos across five deployments, recorded at 1920×1080 resolution (16:9 aspect ratio) with an average data rate of 29.1 ± 3.6 Mbps and an average duration of 11.8 ± 4.5 minutes per video (Figure 3.4, Appendix E.1). The MHI subset comprised 56 videos across five deployments, varying in resolution from 1920×1080 to 3840×2160 (aspect ratios 4:3 and 16:9), with a higher average data rate of 80.6 ± 24.8 Mbps and shorter average video length (7.4 ± 2.8 minutes).

To assess tradeoffs between annotation time and detection accuracy, we employed a three-pronged annotation protocol: (1) a fully automatic method using raw AI detections without manual review: (2) a semi-automatic method that involved post-processing model detections to remove false positives: and (3) a fully manual method, in which the observer reviewed footage at 1× speed and annotated shark occurrences by hand. We assumed the latter method represented the upper bound of detection accuracy (i.e., 100%), under the assumption that a careful manual review would not miss any sharks present in the footage. For model inference, we applied a detection confidence threshold of 0.5, meaning only detections with at least 50% probability were retained. This threshold was chosen to balance sensitivity (detecting as many true sharks as possible) against precision (limiting false positives). We processed videos with a frame rate of 0.2 FPS (one frame every five seconds). To further accelerate object detection and reduce computational load, we resized video frames to a standardized width of 640 pixels. This configuration balances detection coverage and processing efficiency, making it suitable for identifying slow-moving or briefly visible sharks without overwhelming computational or manual review resources. All video processing was performed on a MacBook Pro (2024) equipped with an Apple M4 chip (10-core CPU), 16 GB

of unified memory, and running macOS Sequoia 15.5.

### Performance

We recorded total annotation time for each workflow and quantified performance using two complementary metrics.

First, we calculated overall detection accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

with True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). This metric captures the model's ability to correctly identify both the presence and absence of sharks, offering a broad assessment of detection reliability across video frames. While automatic and semi-automatic methods offer substantial efficiency gains, they are unable to correct false negatives and therefore rely on the fully manual approach to establish a ground-truth benchmark for missed detections. Further, given that we processed one frame every five seconds, any shark that appeared for less than this duration, between the sampled frames, may not have been captured. In such cases, this was deemed a false negative.

We also computed recall, precision, specificity, and the $F_1$ score:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.4}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

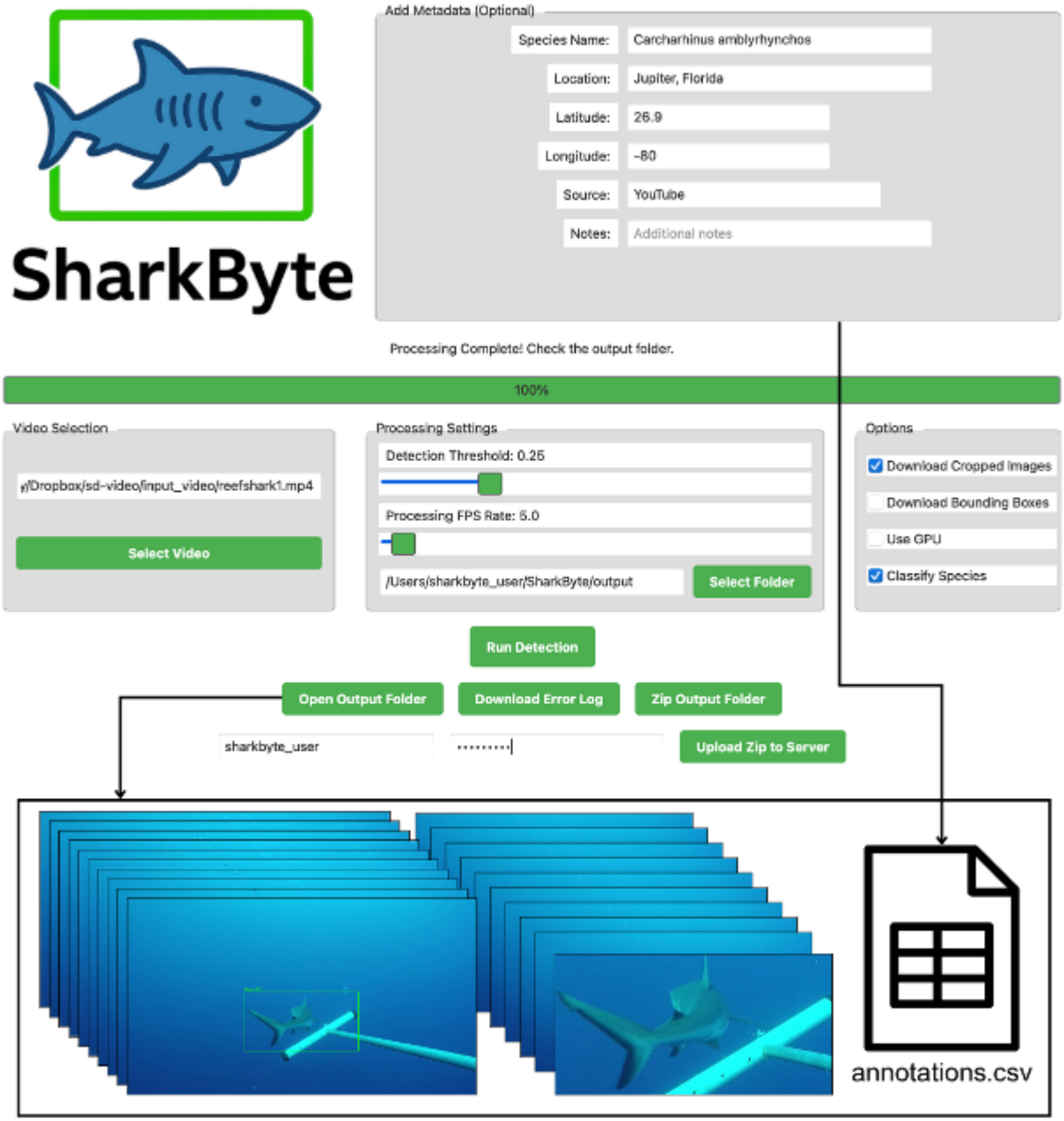


Figure 3.4: The SharkByte GUI enables video-based shark detection and classification.

$F_1$ reflects the balance between precision and recall and focuses exclusively on the model's performance in detecting sharks (the positive class). This is particularly relevant because true positive detections are more useful for evaluating the recall of the SC at the order, family, genus, and species level (Figure 3.1). This comparative framework allowed us to assess both the operational speed and predictive reliability of SharkByte under realistic field conditions.

Finally, we manually calculated species-specific Mean Maximum Number (MaxN) as an index of relative abundance:

$$\text{Mean MaxN} = \frac{\sum \text{MaxN}_h}{\text{Total hours}} \tag{3.6}$$

where $\text{MaxN}_h$ represents the maximum number of individuals of a given species observed in deployment $h$, and values were averaged across deployments in which the species was detected. This metric provides a standardized, non-redundant estimate of relative abundance commonly used in BRUVs biodiversity surveys, minimizing the risk of double-counting individuals [6].

## 3.3 Results

### 3.3.1 Species Classification

Using `sharkDetectoR`, we classified 46,332 held-out sharkPulse images in 16 minutes with the `classify_image` function. The taxonomic classifier SC achieved an end-to-end accuracy of 91.4%, meaning the entire classification pathway (order → family → genus → species) was correctly identified. Boosting and retraining new data pushed the same end-to-end accuracy

to 92.0% (Appendix C). At individual taxonomic levels, accuracy was 96.9% for order, 95.5% for family, 93.2% for genus, and 92.0% for species.

### 3.3.2 SharkByte Performance

Automatic processing took 46.6 minutes for Palau footage and 155 minutes for MHI footage, reflecting increased processing demands of higher-resolution videos. We processed 9,936 frames in total, of which 644 contained sharks: 599 were correctly detected (45 missed), while 370 empty frames were falsely identified as containing a shark. Automatic annotation accuracy was similarly high for both sites, at 95.1% in Palau and 94.3% in the MHIs. Implementing a semi-automatic workflow (manually reviewing and removing false positive detections) marginally increased accuracy by 0.9% in Palau (96.0%), and 5.5% in the MHIs (99.8%) at a modest time cost of 4.6 and 10.3 additional minutes, respectively. Conversely, fully manual annotation, while achieving 100% accuracy, was considerably slower, taking on average 9.2 (Palau) and 2.7 (MHIs) times longer than the automatic workflow, 12.2 and 7.5 hours respectively. SharkByte flagged only 4% of empty frames as containing a shark.

Notably, shark detection performance varied by location and shark density: in the MHI survey, where the MaxN never exceeded one shark per frame (see Table C.1), SharkByte correctly identified sharks in 16 of 22 occupied frames. In Palau, with a higher Mean MaxN of $3.5 \pm 0.05$, SharkByte detected at least one shark in 583 of 622 shark-containing frames but missed some individuals across 248 frames ($n = 1.5 \pm 0.7$ individuals).

Overall, we observed seven shark species across both surveys, including five *Carcharhinus* species, tiger shark (*Galeocerdo cuvier*), and sicklefin lemon shark (*Negaprion acutidens*) (Table C.1). Given a marginal increase to annotation time, the semi-automatic workflow achieved an $F_1$ detection score of 82% and a recall accuracy of 76.5% when classifying cropped

individuals (Figure 3.2).

| | | MHI | | | | | | Palau | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Species** | **Common Name** | **Mean MaxN** | ***n*** | **% Drops** | **Depth (m)** | **$F_1$** | **Recall** | **Mean MaxN** | ***n*** | **% Drops** | **Depth (m)** | **$F_1$** | **Recall** |
| *C. amblyrhynchos* | Grey reef shark | 0.02 ± 0.01 | 2 | 2.7 | 30–34 | 0.95 | 0.71 | 1.4 ± 0.06 | 13 | 27 | 10 | 0.92 | 0.72 |
| *C. falciformis* | Silky shark | — | — | — | — | — | — | 0.15 ± 0.3 | 4 | 9 | 10 | 0.93 | 0.80 |
| *C. galapagensis* | Galapagos shark | 0.01 ± 0.01 | 1 | 1.4 | 30 | 0.5 | 1.0 | — | — | — | — | — | — |
| *C. limbatus* | Blacktip shark | 0.01 ± 0.01 | 1 | 1.4 | 56 | 0.75 | 0.93 | — | — | — | — | — | — |
| *C. plumbeus* | Sandbar shark | 0.01 ± 0.01 | 1 | 1.4 | 45 | 0.35 | 0.63 | — | — | — | — | — | — |
| *G. cuvier* | Tiger shark | — | — | — | — | — | — | 0.04 ± 0.03 | 1 | 9 | 10 | 0.98 | 0.87 |
| *N. acutidens* | Sicklefin lemon shark | — | — | — | — | — | — | 0.04 ± 0.03 | 1 | 9 | 10 | 0.82 | 0.55 |
| **Subtotal all species** | | 0.06 ± 0.01 | 5 | 6.8 | 30–56 | 0.68 | 0.75 | 3.5 ± 0.05 | 19 | 45 | 10 | 0.80 | 0.71 |

Table 3.1: Summary of shark species observed from BRUVs surveys conducted in the MHIs and the Palauan Archipelago. The table reports the average relative abundance (Mean MaxN $hr^{-1}\pm$ SE), total number of individuals observed ($n$), percentage of deployments where each species was observed (% Drops), and depth or depth range of deployments (Depth in meters). Detection performance is evaluated using the $F_1$ score, reflecting the accuracy of automatically detecting sharks within video frames, and recall, representing the accuracy of subsequent boosted species-specific classification following the data augmentation protocol. Notably, *Carcharhinus amblyrhynchos* (grey reef shark) was the most frequently observed species in Palau (27% of deployments), with notably higher average abundance (Mean MaxN = 1.4 ± 0.06) and an increased classification recall of +17% (after data augmentation), compared to the MHIs (2.7% of deployments: Mean MaxN = 0.02 ± 0.01, Δ Recall = +5.0%).

### 3.3.3 Processing and Annotation Performance

Several factors influenced processing and annotation performance. Higher-resolution and higher data-rate videos increased processing times by approximately 3.3× per minute of footage (MHI vs Palau) but improved the model's ability to capture fine morphological features. At a low sampling rate (0.2 fps), our detector consistently flagged the aggregations of shark activity, so every shark individual that appeared for a typical residency was captured at least once, even though many shark-containing raw frames were skipped. This strategy minimized processing time while preserving sensitivity to individuals, with the only misses being very brief (<5 seconds), distant in the frame, and isolated from any other shark-rich segments.

Detection performance was also influenced by background complexity, lighting, and shark behavior. In Palau, pelagic BRUVs deployed at 10 meters depth benefited from higher ambient light levels and simpler backgrounds, improving the visibility of distinct morphological features. Conversely, benthic BRUVs in the BRUVs recorded footage with complex backgrounds (algae, coral reefs, rock formations), high particulate activity, and lower light conditions, contributing to both false positives (e.g., bait canisters, small fish, crabs, vegetation) and occasional false negatives. While no sharks were completely missed in the MHI footage, Palau experienced a higher average duration of shark presence per video that contained at least one shark (7.8 minutes longer) and shark abundance (Mean MaxN 5.8 times higher), challenging automatic detection, particularly when sharks lingered in distant backgrounds.

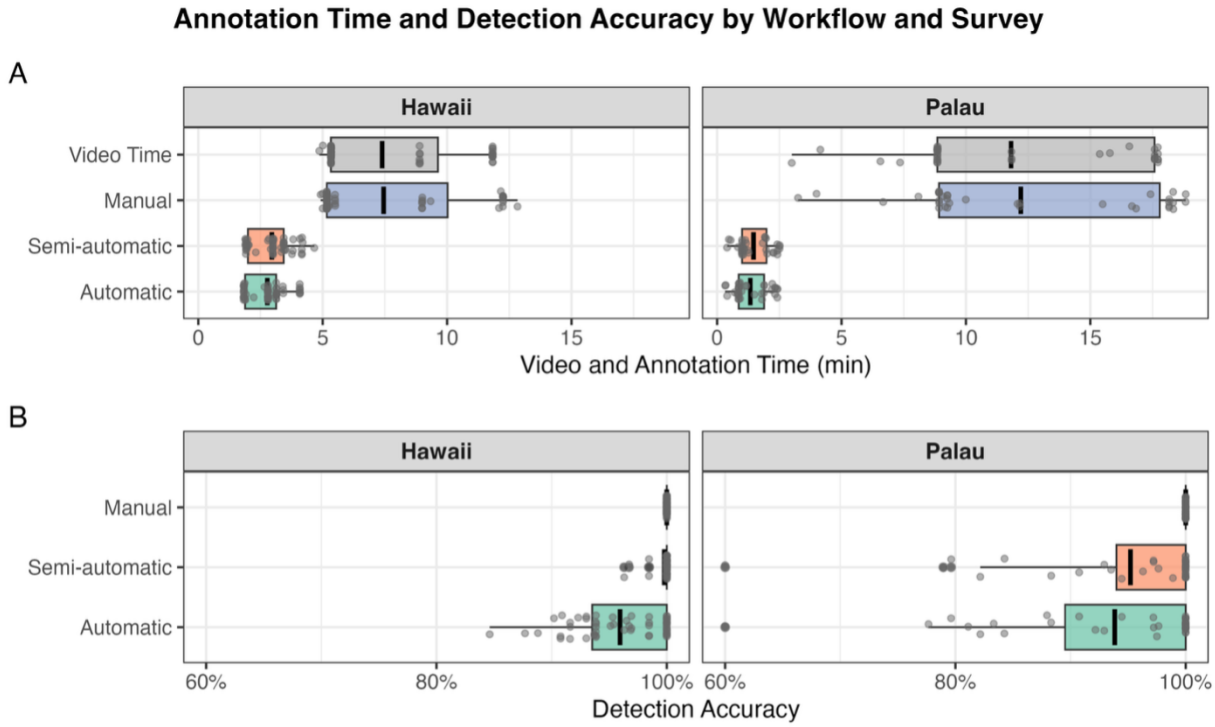


Figure 3.5: Annotation time and detection accuracy is represented across processing workflows and surveys. Panel (A) shows the time required to annotate shark detections in videos across the three workflows. Times are shown alongside raw video duration for comparison. Panel (B) shows detection accuracy across workflows. Boxplots summarize distributions with points representing individual videos and vertical bars indicating mean values.

## 3.4 Discussion

BRUVs and other non-lethal survey approaches are increasingly used as cost-effective tools to monitor shark populations and address data gaps [60]. However, a persistent limitation is the substantial effort required to manually process this material [167], which often involves reviewing thousands of hours of footage and can strain both time and resources. To address this challenge, we developed an analytical pipeline that automates postprocessing through a hierarchical CNN, adapted from our previous classifier [78], and a YOLO-based detection module optimized for underwater shark footage [164]. These were packaged into two accessible, cross-platform tools: the programmatic `sharkDetectoR` package and the SharkByte application. We evaluated their performance by processing 46,332 shark images and 13.9 hours of BRUVs video, achieving robust detection and classification accuracy while substantially reducing annotation time. This workflow further integrates with the sharkPulse platform, enabling streamlined data submission, expert validation, and retraining to iteratively improve species-specific ecological data and model performance.

The SC demonstrated strong classification performance, achieving 92% accuracy across previously unseen datasets of 46,332 sharkPulse images [52]. Crucially, the hierarchical CNN structure strategically enhanced accuracy by sequentially conditioning predictions at each taxonomic level (order $\rightarrow$ family $\rightarrow$ genus $\rightarrow$ species), thereby masking unlikely taxa and reducing misclassification among visually similar taxa. Automatic classification was made quicker by employing the GPU-accelerated API processing functions that are directly accessible through the `sharkDetectoR` package. This processing architecture facilitated seamless integration into ecological workflows, significantly broadening usability and scalability, enabling more reliable species-specific annotations across diverse platforms and heterogenous datasets.

In processing BRUV surveys, overall automatic annotation accuracy was high at 94.7%. Higher-resolution MHI videos resulted in processing speeds roughly three times slower per minute of video than Palau footage. Despite this, detection rates were somewhat lower in MHI, with sharks detected in 72.7% (16 of 22) of shark-containing frames versus 93.7% (583 of 622) in Palau. Notably, Palau videos also featured sharks present for longer durations on average (7.8 minutes longer) and higher abundance (Mean MaxN 5.8 times greater), factors that both facilitated detections yet sometimes challenged the model when multiple individuals appeared at a distance. Such conditions increased FN rates. By consistently detecting the shark nearest the bait, SharkByte enabled reviewers to spot FNs through semi-automatic annotation. In rare cases where sharks were only present at a distance without approaching the bait canister (6.3% of shark frames in Palau), all sharks were missed and could not be recovered even with semi-automatic review, suggesting a need to optimize the confidence threshold further.

Factors such as simpler backgrounds and lighting, and sustained shark presence coupled with processing rate likely influenced detection success in Palau [84, 97, 130]. Moderate reductions in resolution also accelerated processing with minimal loss to detection outcomes, offering a practical tradeoff that decreases manual review demands while maintaining reliable annotation across different logistical and environmental contexts [164]. Where brief appearances or subtle morphological features are critical in the MHI survey, higher video quality may still be advantageous [22], a point future work should explore more systematically.

The broader implications of this work extend beyond methodological innovation, as the tools developed here open new pathways for how sharks and other data-poor taxa can be monitored in practice. By automating time-intensive tasks, the SD allow managers to process thousands of hours of BRUV or opportunistic image and video data that would otherwise be infeasible to review manually. This capability translates directly into practical appli-

cations: for example, fisheries observer programs can integrate the workflow to flag shark bycatch in near real-time, while marine protected area managers can use automated video annotation from BRUVs or dockside cameras to generate standardized records of shark encounters. Port-based footage can directly verify compliance with no-take regulations, while BRUV biodiversity and abundance baselines track ecological outcomes of protection over time, providing a dual tool for enforcement and adaptive management. The system also empowers fishers and citizen scientists by returning immediate identifications and logbook-ready annotations that incentivize data sharing while reducing barriers to participation. Permitting and licensing at the commercial and recreational levels could also be linked to automated bycatch identification systems, significantly expanding the reach of shark monitoring with species-specific indices. Taken together, these applications illustrate how these tools can deliver actionable information to diverse stakeholders, supporting both top-down policy decisions and bottom-up community engagement in shark conservation.

We highlighted the potential for enhancing AI-driven ecological monitoring through expanded citizen scientist engagement, accessible automated tools, and a robust cyberinfrastructure (sharkPulse). The groundwork established here is streamlined automatic annotation workflows, positioning sharkPulse, the SD, and integrated open-source tools as flexible, robust, and continuously improving platforms. By scaling data generation and model sophistication in parallel, these efforts advance the frontier of automated ecological monitoring, helping to close critical knowledge gaps and providing conservation and management with a powerful, adaptive toolkit for safeguarding sharks.

# Chapter 4

# Leveraging Social Networks and Open Data for Inferring Shark Population Trends

# Abstract

Social networks (SNs) and open biodiversity platforms now generate unprecedented volumes of geotagged media documenting human–wildlife encounters worldwide. Harnessing these data for ecological inference remains challenging due to platform heterogeneity, inconsistent metadata, and strong human-driven biases. Here, we develop and benchmark a reproducible workflow that converts raw posts into standardized ecological observations and relative abundance indices for sharks. We integrate computer vision for automated detection and taxonomic labeling, natural language processing for spatiotemporal metadata extraction, and platform-specific filtering to address duplication, aquarium records, and inland geotags. We apply this data-crowdsourcing and sanitizing framework to four major data sources: Instagram (IG), Flickr, iNaturalist (iNat), and Global Biodiversity Information Facility (GBIF). We then predict relative abundance as Sightings per Unit Effort (SPUE) using a negative binomial likelihood with user-activity offsets to control for observation effort. To evaluate predictive performance, we compared SN trends with independent long-term Catch Per Unit Effort (CPUE) and Baited Remote Underwater Video (BRUV) records from well-studied regions, in the Bahamas and Hawaii, that serve as benchmarks for shark population research. INat records produced the most reliable SPUE trends, while SN data (IG, Flickr) delivered unmatched volumes of raw shark observations. Modeled trajectories recovered broad regional patterns observed in conventional monitoring, including multispecies increases in the Bahamas and reef-shark declines around the Main Hawaiian Islands (MHIs). Together, these results demonstrate that opportunistic digital observations can yield indicators of relative abundance when standardized for effort and bias. Beyond case studies, this framework establishes a scalable blueprint for integrating heterogeneous digital data into wildlife monitoring.

## 4.1 Introduction

Sharks are important ecological drivers and charismatic predators that help maintain the balance of marine environments [40]. Yet, they face increasing threats from fishing pressure, habitat degradation, and poorly informed conservation and management. These factors have contributed to a 71% decline in global abundance since 1970 [117]. Observation data from standardized surveys and fisheries monitoring remain costly and logistically challenging to collect, and are often undervalued compared to higher-profile fisheries such as tunas and billfish [43]. As a result, sharks are among the most data-deficient marine groups, with limited species-specific abundance and distribution indices available for driving actionable policy [31, 44].

Social networks (SNs) and open biodiversity platforms host immense repositories of visual content that capture human–wildlife encounters across broad spatial and temporal scales [108]. These posts and records represent an underused source of ecological information [64]. Instagram (IG), for example, is among the largest image-sharing platforms globally, with billions of daily uploads: as of September 2025, more than 6.7 million posts include the hashtag `#shark`. Flickr, with >10 billion archived photographs, caters to photography enthusiasts with rich metadata (e.g., capture time/location, camera settings). Open biodiversity platforms such as iNaturalist (iNat), eBird, and Global Biodiversity Information Facility (GBIF) provide vetted species observations: iNat includes >153k and ray records spanning 1,251 species, and GBIF hosts >600k shark occurrences with links to survey programs. Together, these sources deliver opportunistic, global coverage that is difficult to achieve with traditional monitoring alone [108, 159].

Geotags and timestamps enable precise locations of encounters, supporting analyses of occurrence, relative abundance, migratory and seasonal activity, and shifts in distribution

ranges [149]. Hashtags function as simple classification filters that group posts into relevant data pools (e.g., `#shark`, `#greatwhiteshark`), allowing targeted retrieval of shark-related content. Captions, comments, and tags can enhance the reliability of taxonomic and spatiotemporal information [127]. Yet harnessing these data for population inference requires addressing platform-specific biases and enforcing reproducible standards.

A central challenge for shark conservation is the scarcity of reliable, species-specific population indices, especially in regions where formal monitoring is limited. sharkPulse demonstrated that SNs can generate verified records of global shark presence [52]. However, social media is not typically designed for ecological monitoring: posts cluster around popular dive sites, tourism seasons, or fishing activity, and their quality varies depending on the user and the platform. Incorporating transparent user activity is the next crucial step in standardizing SN records so that indices reflect biological abundance rather than fluctuations in how many people are posting.

Different platforms pose different challenges for meeting metadata quality requirements and revealing user engagement. IG contains unmatched volumes of shark encounters but severely restricts access to data and user activity through its official Application Programming Interface (API) [12]. Flickr provides spatiotemporal indicators but has more limited user engagement. By contrast, iNat and GBIF supply higher-quality identifications and transparent access policies, though their coverage is narrower in scope. These platform differences induce distinct biases that must be acknowledged when interpreting opportunistic records. *Visibility bias* favors charismatic species and popular dive sites, *participation bias* reflects patterns in human population density and tourism, and *annotation bias* arises from variable taxonomic skill and incomplete metadata [160]. Left unaddressed, these biases risk producing indices that reflect human behavior without a clear indication of ecological dynamics.

Overcoming these challenges requires systematic workflows that can filter noisy content, standardize observations against measures of posting effort, and provide explicit estimates of uncertainty. In this way, opportunistic posts can be transformed into ecological indices more comparable to conventional surveys. Our study builds on prior work showing that SNs can yield valid species occurrences [52], extending this potential to evaluate population trends while critically examining the limitations and biases unique to each platform.

In this study, we evaluate these platforms across the dimensions of access, metadata quality, and observation volume (Figure 4.1). We assess how SN records can be converted into meaningful indicators of shark population change. We leverage IG, Flickr, iNat, and GBIF as complementary sources of shark occurrences and develop semi-automated, reproducible pipelines to source, filter, and annotate posts for population inference. We focus on two well-studied regions (the Bahamian and Hawaiian archipelagos) to benchmark trends as Sightings per Unit Effort (SPUE) trajectories. We compare these trends with historical assessments in the region that have generated independent indices (e.g., Baited Remote Underwater Video Systems (BRUVs) and fishing surveys). Our aim is not to replace conventional monitoring, but to provide a transparent blueprint for integrating large, opportunistic digital signals with formal ecological assessments. This approach expands spatial coverage, increases temporal resolution, and reduces the cost of generating species-specific indices while explicitly accounting for bias and uncertainty.

## 4.2 Methods

We evaluated several systematic approaches for accessing and processing SN data, with particular attention to platform limitations, ethical and legal considerations, and data quality challenges (Figure 4.2). When species-specific filtering options were unavailable, we fo-

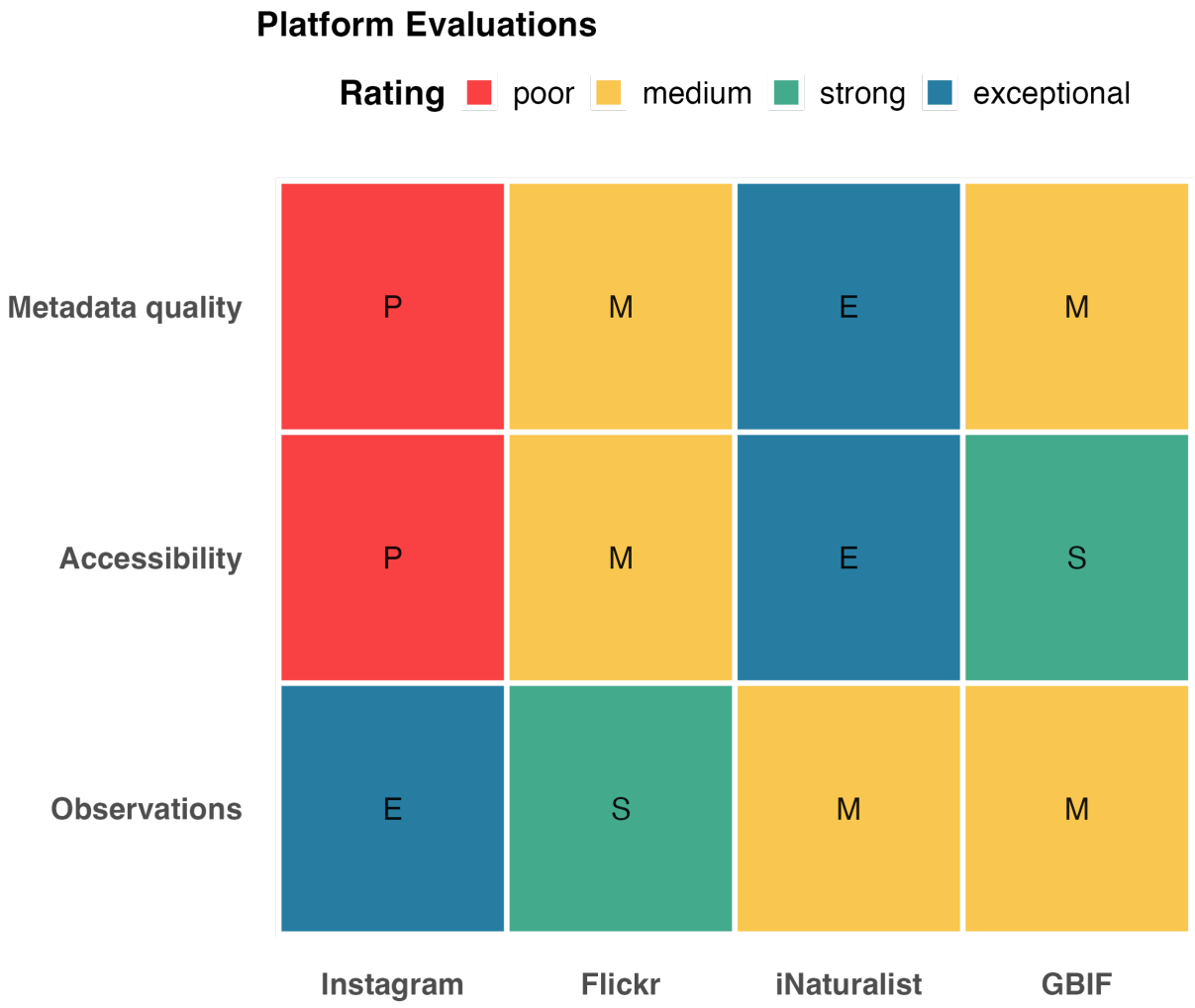


Figure 4.1: Evaluation of raw observation potential, accessibility, and metadata quality by platform.

cused on distinct shark hashtags or textual searches, employing both open-source tools and proprietary third-party platforms to collect posts. Each approach had unique operational, ethical, and accessibility constraints that influenced the comprehensiveness and transparency of data collection. These limitations highlighted the need to account for inaccessible posts, duplicated content, and unknown retrieval constraints. We applied several filtering and annotation workflows designed to automatically generate shark observations with crucial ecological metadata such as location and time. This ecological context was assigned using user- and platform-generated metadata as well as morphological attributes automatically classified from images. Subsequently, depending on the platform, we sourced sub-samples or total user engagement to standardize observations.

To evaluate relative shark abundance, we converted posts into effort-standardized indices for two data-rich regions—the Bahamas and Hawaii—where long-term monitoring and historical surveys provide useful context. Records were aggregated into monthly units and paired with platform-specific proxies of observation effort. For each temporal bin, we computed SPUE indices and predicted temporal patterns using Generalized Linear Models (GLMs), assuming counts followed a negative binomial distribution. Both human-validated and automatically identified shark observations were analyzed to assess trade-offs between manual versus automated pipelines. Resulting indices were compared with independent survey trends to evaluate the effectiveness of our data workflows and modeling choices for guiding future efforts.

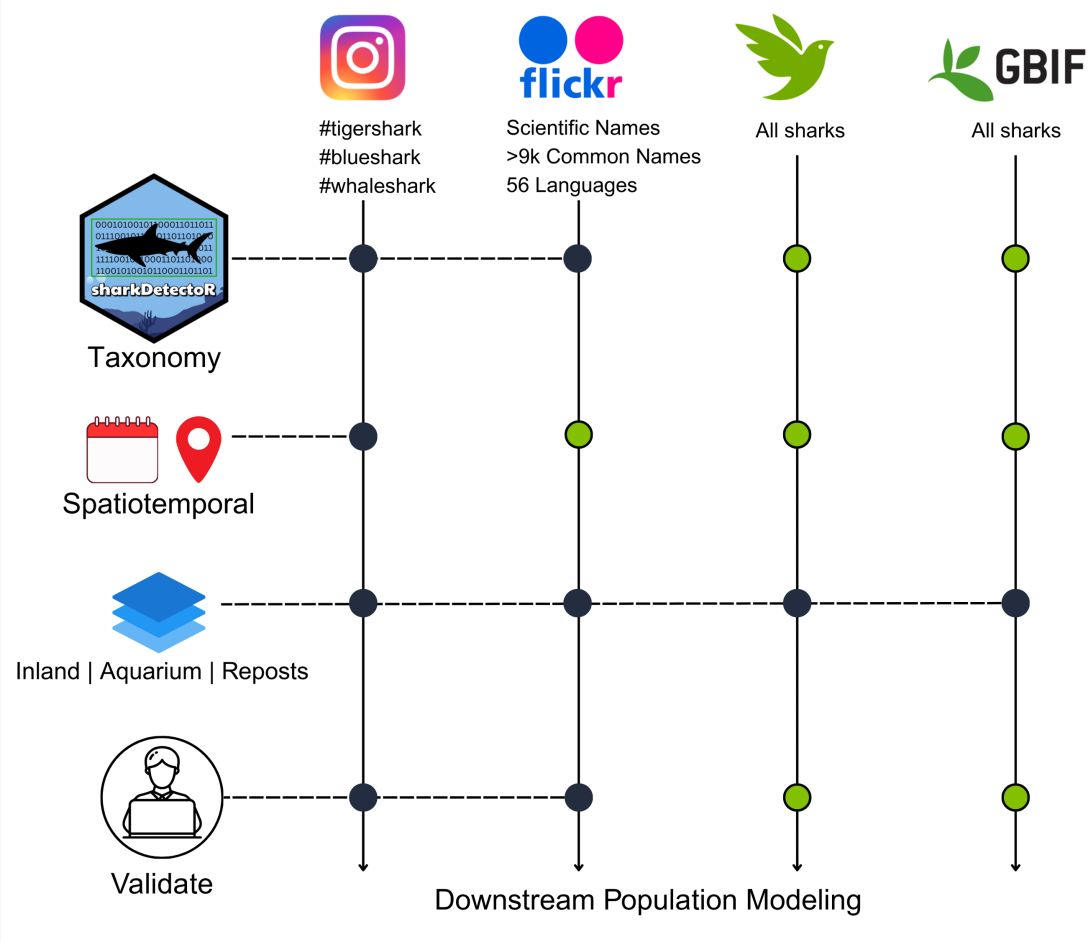


Figure 4.2: Filtering and data-assignment workflows by platform. The black dots are filtering steps, and the green dots represent metadata that is already available on the platform.

### 4.2.1 Instagram and Flickr

We mined IG posts and associated metadata from three species-specific hashtags: `#whaleshark`, `#blueshark`, and #tigershark (Figure 4.2). These species were selected because they are morphologically distinct, rich with data, rarely confused with other taxa, and therefore simpler to validate. We collected posts with two open-source tools: `InstaCrawlR` [132] and a third-party scraping platform, Apify [155].

From Flickr, we obtained posts through the public API. We adapted the `photosearcher`

package [55] to automatically subdivide the retreival date windows whenever Flickr's 4,000-record per-query limit was reached. Queries were built from a PostgreSQL table `shark_names` containing 9,453 common names corresponding to 532 shark species in 56 languages, compiled from FishBase [17, 57]. Searches were performed separately for common and scientific names. Only posts with geolocation metadata were retained, including both date-taken and date-uploaded fields. Previously completed queries were tracked to avoid redundancy, and randomized pauses were inserted to comply with API rate limits.

After posts were obtained, we assigned taxonomic labels with the hierarchical Shark Detector (SD) framework [52, 78] (Chapter 2, 3). This model applies a binary shark vs. non-shark classifier, followed by conditional classifiers at the order, family, genus, and species levels. The training dataset, sourced from sharkPulse, comprises over 200k images of 80 shark species [52]. Model training and application were implemented using the `sharkDetectoR` package [75] (Appendix C).

After assigning taxonomic labels, we assigned location to IG posts because geotags are stripped upon upload. IG posts were geolocated using the Google Maps API geocoding function [62] applied to the post text metadata. Where explicit location tags were available (e.g., Tiger Beach, Bahamas), we queried the geocoding function and stored the centroid of the returned place polygon as the post coordinates. For posts without explicit location tags, the full post text was submitted to the same API, and the first set of returned coordinates was recorded. We validated the accuracy of this approach by surveying 197 IG users of posts that were automatically identified as a shark, assigned location and date, and then validated by a human. We asked the user to confirm the precise location and date of the sighting. To remove duplicate IG posts, we used the `hashlib` Python library [66] to compared pixel values across the pool of geocoded images. This allowed us to identify and flag reposts of the same shark observation. In addition to geocoding, we systematically filtered out misclassified or

non-wild observations. Posts falling within 10 km of inland areas were flagged as erroneous geotags, while posts within 1 km of known aquaria were classified as aquarium records rather than wild observations [52].

We manually reviewed IG and Flickr posts to assess the accuracy of automatic data-generation approaches. To scale this process, we developed a dedicated web application called the Validation Monitor which compiles potential shark posts into a streamlined interface where citizen scientists can confirm species identity, location accuracy, and whether an observation occurred in an aquarium, while also adding notes. Upon a citizen science validation, a sharkPulse expert revalidated the sighting to append a higher confidence label and assess the precision of automatic taxonomic labeling. Citizen scientists reviewed 4,844 IG posts and 5,679 Flickr posts in total. At least one sharkPulse expert revalidated 4,543 IG posts and 933 Flickr posts that were previously labeled by the SD and then a citizen scientist. With the revalidation checks, we calculated the species-level classification accuracy of the SD to correctly identify posts from both SNs.

### 4.2.2 iNaturalist and GBIF

To obatain appropriate records from iNat, all shark observations were queried using the `rinat` R package [11]. A one-time global sweep collected all shark records, which were archived in a PostgreSQL table. Observations were downloaded with filters for taxon, date, and geographic bounds. Monthly automated updates retrieve only new records.

Global shark records were retrieved from GBIF using bulk-download services targeting Selachimorpha with the `rgbif` R package [29]. Metadata were downloaded in a Darwin Core (DwC) archive format [172], including linked still images. Each record was linked to event- and parent-event identifiers, allowing survey-level grouping. To avoid duplication

between the iNat scrape and iNat records ingested by GBIF, we excluded iNat's dataset key and parsed identifiers embedded in GBIF fields. Records were inserted into a unique PostgreSQL table and curated for valid coordinates, dates, and taxonomy.

We accepted `Research Grade`-labeled records as ground truth because they require date, location, media, and community taxonomic consensus (two-thirds agreement among identifiers), and form the basis for data shared between iNat and GBIF. The remaining records were labeled as `NeedsID` or `Casual` indicating a taxonomic consensus had not been reached, thus we excluded these from downstream analyses. Due to the records' verified taxonomy and high quality metadata, we did not need to employ the `sharkDetectoR`, spatiotemporal assignment workflows, or post-crowdsourcing human checks (Figure 4.2). However, we filtered posts geolocated inland and near or within an aquarium to retain more precise spatiotemporal records.

### 4.2.3 Observation Effort: Flickr and iNaturalist

Quantifying observation effort is essential for interpreting changes in shark reporting activity and distinguishing true biological trends from fluctuations in user engagement. Because social and citizen science platforms differ widely in accessibility and data structure, direct measures of effort must be adapted to each source. This section describes how effort was estimated or inferred for platforms with openly accessible metadata, focusing on Flickr and iNat where reliable indicators of user activity could be derived.

Generating transparent measures of IG user activity—or even reliable subsamples as proxies for observation effort—was not feasible due to strict privacy protections and limited data accessibility. In contrast, GBIF provides meaningful indicators of observation effort: however, shark records are aggregated across numerous individual surveys that report effort

using inconsistent formats and indices. As a result, additional data-structuring workflows are required to harmonize and integrate these heterogeneous survey strategies. From both platforms, however, verified shark observations were immediately piped into the training dataset of the SD, increasing its predictive performance.

We quantified user effort associated with shark observations on Flickr and iNat to standardize sightings across space and time. For Flickr, we treated the density of non-shark images (classified from the broader pool of platform posts) as a proxy for user activity, pairing these records with shark observations within the same spatiotemporal bins. For iNat, we estimated effort more directly by retrieving all non-shark observations recorded within the same spatial and temporal boundaries as the shark queries, thereby providing an index of overall reporting intensity.

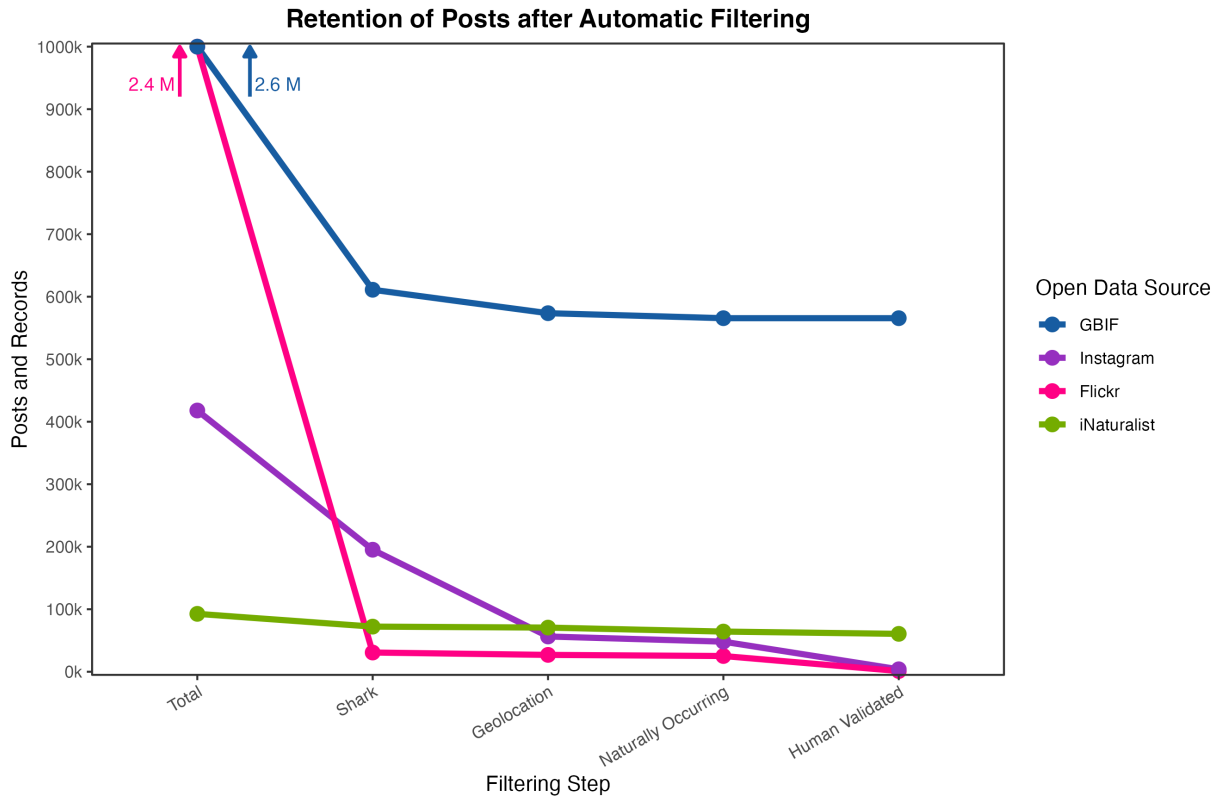


Figure 4.3: Automated workflow for sourcing and filtering shark observations from four major open data platforms. Total posts represent the first ingested pool of images, records, and/or posts from each platform.

### 4.2.4 Predicting Relative Abundance

To model temporal patterns in shark observations, we compiled the naturally occurring, georeferenced posts within the study-region bounds of the Main Hawaiian Islands (MHIs) and the Bahamian archipelago. Code to reproduce these methods can be found in Appendix A. Two models were fitted to the number of shark observations ($Y_t$) per month (Equation 4.2) or year (Equation 4.3), with corresponding measures of user effort ($E_t$) expressed as the total number of posts or observations on the same platform. Because these data arise from opportunistic reporting, counts of shark posts are typically overdispersed relative to a Poisson

process. We therefore modeled $Y_t$ using a negative binomial distribution (Equation 4.1) with mean $\mu_t$ and dispersion parameter $k$:

$$Y_t \sim \text{NegBin}(\mu_t, k), \quad \log(\mu_t) = \eta_t \tag{4.1}$$

where $\eta_t$ is the linear predictor. User effort ($E_t$) was incorporated as an offset term to standardize for unequal reporting intensity across time. Models 1 and 2 were plotted on the same graph as a trend line and yearly predictions respectively, illustrating the results in Figures 4.8-4.11.

**Model 1: Long-term Smooth Trend.** To capture smooth, gradual changes in relative shark abundance, we first modeled time as a continuous covariate:

$$\eta_t = \beta_0 + \beta_1 \left(\text{Year}_t\right) + \gamma_1 \sin\left(\frac{2\pi \,\text{Month}_t}{12}\right) + \gamma_2 \cos\left(\frac{2\pi \,\text{Month}_t}{12}\right) + \log(E_t) \tag{4.2}$$

Here, the sine and cosine terms explicitly model cyclic seasonal effects within a calendar year, ensuring reproducibility of the seasonal component. The coefficient $\beta_1$ describes the long-term temporal trajectory.

**Model 2: Independent Yearly Estimates.** To evaluate year-to-year variability independently of any assumed functional form, we fitted a parallel model treating year as a categorical factor:

$$\eta_t = \alpha_0 + \alpha_{\text{Year}(t)} + \gamma_1 \sin\left(\frac{2\pi \,\text{Month}_t}{12}\right) + \gamma_2 \cos\left(\frac{2\pi \,\text{Month}_t}{12}\right) + \log(E_t) \tag{4.3}$$

where $\alpha_{\text{Year}(t)}$ denotes the coefficient associated with Year $t$, where each year is treated as a categorical factor. This formulation yields discrete, independent point estimates of expected

shark counts per year.

**Deriving Relative Abundance Indices.** From both models, the fitted mean counts $\hat{\mu}_t$ were converted to SPUE indices by standardizing to a fixed unit of user effort (1,000 posts or records):

$$\text{SPUE}_t = \frac{\hat{\mu}_t}{E_t/1000} \tag{4.4}$$

This standardization removes the effect of changing platform activity, allowing temporal comparison of relative abundance across years and between platforms. Model 1 provides a smoothed representation of overall temporal trends, while the independent yearly model highlights interannual fluctuations and uncertainty. Goodness-of-fit diagnostics and residual plots evaluating model performance are provided in Appendix G.

### 4.2.5 Comparative Analysis

To evaluate whether SN abundance indices captured the direction of known population trends, we compared our SPUE time series with independent assessments of shark abundance. The comparison focused on taxa well represented in our datasets and with documented monitoring records from overlapping or ecologically adjacent regions.

We expanded our spatial scope beyond the immediate archipelagos (Hawaiian and Bahamian) to include relevant studies from the broader Central Pacific and Western Atlantic basins, as population trajectories for many shark species are assessed at regional or oceanic scales.

**Data sources and search strategy.** We performed targeted searches across scholarly databases (*Google Scholar*, *Web of Science*) and institutional portals (e.g., International

Union for Conservation of Nature (IUCN) Red List species accounts, national and state resource agencies, longline observer programs, and regional field stations). Search strings combined regional terms (e.g., "Hawaii," "Northwestern Hawaiian Islands," "Bahamas," "Bimini," "Eleuthera," "Caribbean," "Central Pacific") with taxonomic and methodological keywords ("shark*", "abundance", "trend", "Catch Per Unit Effort (CPUE)", "BPUE", "biomass", "catch rate", "Maximum Number (MaxN)", "encounter rate", "longline", "BRUV", "diver survey").

Eligible sources were required to report multi-year abundance indices or clear trend statements at the species or assemblage level, with sampling effort and methods explicitly defined. Spatial footprints were checked to ensure coverage within or directly adjacent to our focal regions.

**Data extraction.** For each source, we recorded the following: region or site(s), species or species group, study period, index type (e.g., CPUE, BPUE, biomass, MaxN, encounter rate), and the direction of the reported trend (increase, decrease, stable, mixed/uncertain). When raw time-series data were available, we verified the direction empirically: otherwise, we used the authors' qualitative descriptions. We noted contextual factors that might influence comparability, such as baiting practices, tourism intensity, site fidelity, or habitat differences.

**Comparison and rationale.** We compared the *direction* of change (↑ increase, ↓ decrease, ≈ stable, ? uncertain). This approach reduced biases caused by differences in sampling design, spatial coverage, and survey effort among studies. Focusing on directionality provided a consistent basis for testing whether social media trends aligned with established empirical trajectories.

When multiple species within a region were available, species-level directions were aver-

aged qualitatively to derive a composite multi-species index (Table 4.1). We then classified the level of agreement between iNat SPUE trends and independent sources as:

- **Agree**: same directional trend during overlapping years,
- **Disagree**: opposite direction,
- **Mixed**: conflicting evidence among external sources,
- **Unknown (?):** insufficient reference information.

Agreement between Flickr SPUE and reference trends can be inferred from Table 4.1, but not explicitly reported, as Flickr data include unverified automatic classifications that reduce trend reliability. All comparisons emphasized qualitative concordance in trend direction and temporal consistency across methods.

## 4.3 Results

### 4.3.1 Overview of Social and Open-Data Sources

Across platforms, we compiled a total of more than 5.4 million posts and records from IG, Flickr, iNat, and GBIF. Roughly two million GBIF entries were immediately excluded as absence records from systematic surveys or as non-human observations. Together, the remaining datasets represent one of the largest consolidated records of shark occurrence derived from opportunistic and open-access sources to date. Open-data repositories such as iNat and GBIF provided structured access and consistent metadata that enabled large-scale, reproducible data harvesting. In contrast, SNs such as IG and Flickr required customized data-mining workflows to overcome platform-specific limitations (Figure 4.2), yet ultimately

provided complementary and otherwise unavailable observations (Figure 4.3). These combined sources significantly expanded global coverage of shark occurrences (Figures 4.4, 4.5, 4.6, and 4.7), underscoring the value of integrating citizen science platforms with SN media to reconstruct species distributions at scale.

### 4.3.2 Instagram and Flickr

From IG, we mined 418,091 posts tagged with three species-specific hashtags: `#whaleshark` (*Rhincodon typus*), `#tigershark` (*Galeocerdo cuvier*), and `#blueshark` (*Prionace glauca*). Subsequent data acquisition with `Apify` targeted the hashtags `#tigershark` and `#whaleshark` to broaden coverage and test platform limits, retrieving 85,836 `#tigershark` and 325,110 `#whaleshark` posts between June 2012 and November 2022.

Taxonomic classification using the hierarchical SD framework identified 208,924 posts as containing sharks. Among these, 60,155 posts (29%) were successfully geocoded using the Google Maps API. From surveing the 197 IG users, 99 responded and confirmed timestamp accuracy of 94.9% (within one month) and location accuracy of 82.8% (within 300 km) [52]. Further spatial filtering excluded duplicate and misclassified records, as well as posts occurring within 10 km of land or within 1 km of known aquaria.

From Flickr, we collected 2,356,977 posts spanning 11 September 2001 to 10 August 2025 (Figure 4.3). Posts were obtained through the public API using an adapted version of the `photosearcher` package, which automatically subdivided retrieval windows to bypass the platform's 4,000-record query limit. The search dictionary comprised 9,453 common names corresponding to 532 shark species across 56 languages, compiled from FishBase. Automated monthly updates ensured continuous ingestion of newly uploaded media containing shark names. Taxonomic labeling through the hierarchical SD model identified 51,285 posts as

shark-containing, while the Validation Monitor facilitated manual review of 5,679 posts and expert revalidation of 933 entries. This process yielded a reproducible, scalable mechanism for harvesting and verifying global shark observations from Flickr [52].

#### Automatic Taxonomic Classification

Automatic species-level classification assigned 79 unique species labels to shark posts processed with SD versions 1–4 [52, 78] (Chapters 2, 3). Within the automatically identified pool, 33,835 IG posts were recognized as reposts of the same shark observation. Cross-validation of 4,543 IG posts representing 15 species—each sequentially classified by the binary and multi-class SD, validated by a citizen scientist, and subsequently revalidated by a sharkPulse expert—showed that 3,841 posts (84.5%) were correctly identified to the species level across eight species. Only one post corresponded to a species not currently included in the SD taxonomy (*Carcharhinus perezii*). For Flickr, 933 posts encompassing 40 species met the same validation criteria, of which 681 (73%) were correctly classified across 23 species, while 25 posts (10 species) corresponded to taxa not yet supported by the SD. These results indicate that expanding the SD's taxonomic coverage would further improve classification accuracy, particularly for species more frequently reported on Flickr.

### 4.3.3 iNaturalist and GBIF

As of August 2025, the iNat-sourced dataset contained 57,620 global shark observations spanning from 1920 to the present. Of these, 50,804 records (88%) were classified as `Research Grade` and used as the primary input for abundance modeling.

The curated GBIF dataset comprised 2,635,400 records, of which 625,568 passed all filters for valid coordinates, accepted taxonomy, and human verification. These records were

distributed across 76,929 event identifiers (single surveys) and 4,711 parent-event identifiers (survey groups). Historical records extend back to the 1600s, while the densest concentration of modern observations spans 1945 to July 2025 (assessed August 2025). All GBIF records were cross-checked to remove duplicates with iNat, and inland or aquarium-proximate records were filtered out to retain only reliable, wild observations.

### 4.3.4 Observation Effort: Flickr and iNaturalist

In the Bahamas, we automatically identified 3,360 Flickr shark observations and 23,817 non-shark observations. iNat users submitted 1,102 shark and 57,995 non-shark observations within the study area. In the MHIs, we automatically identified 852 shark Flickr observations and 12,393 non-shark observations. iNat users submitted 1,140 shark and 104,866 non-shark observations. These non-shark records served as a proxy for observation effort.

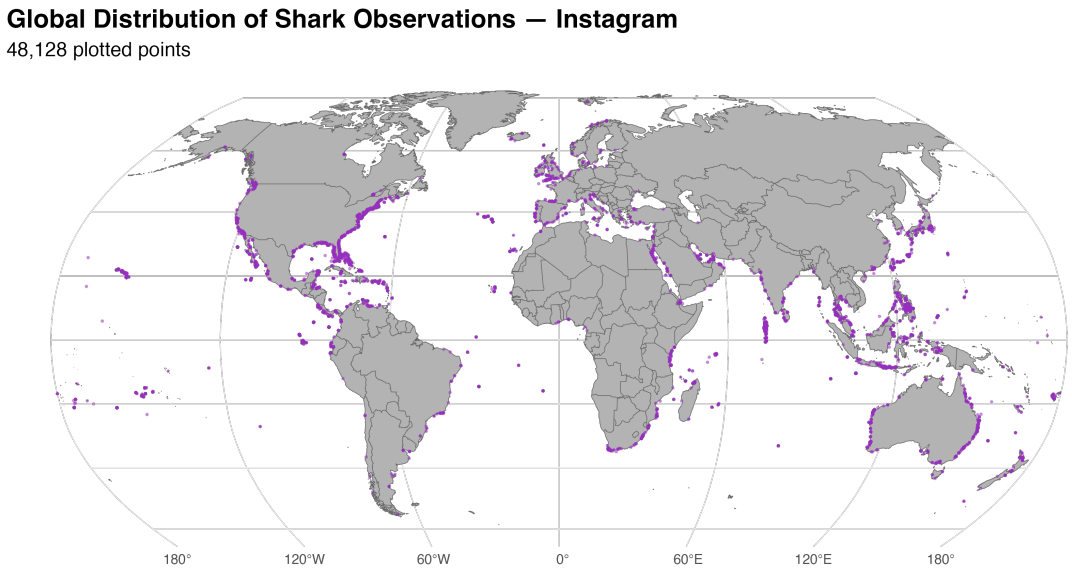


Figure 4.4: Global distribution of IG shark observations.

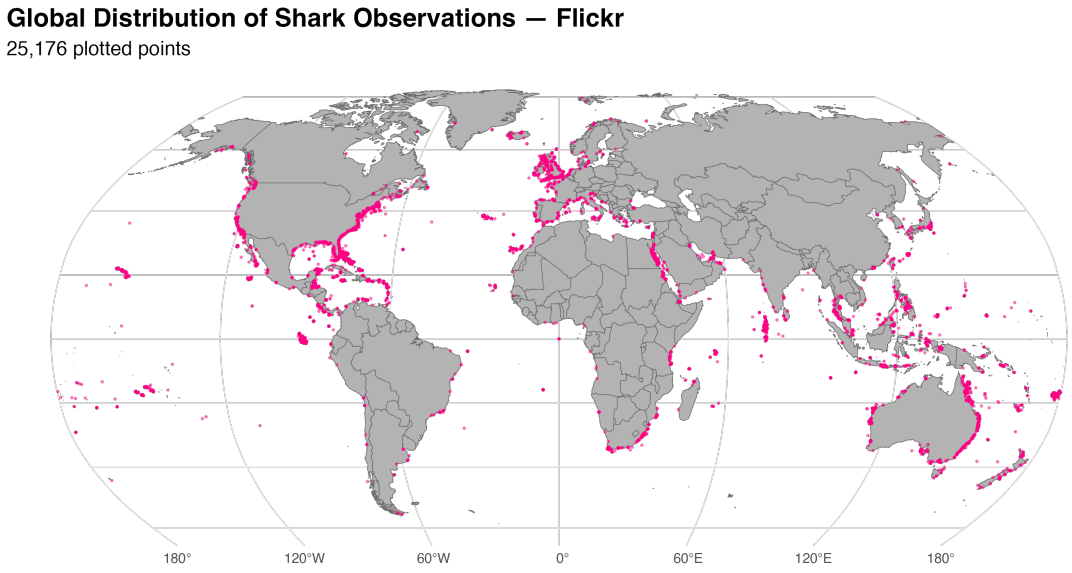


Figure 4.5: Global distribution of Flickr shark observations.

Figure 4.6: Global distribution of iNat shark observations.

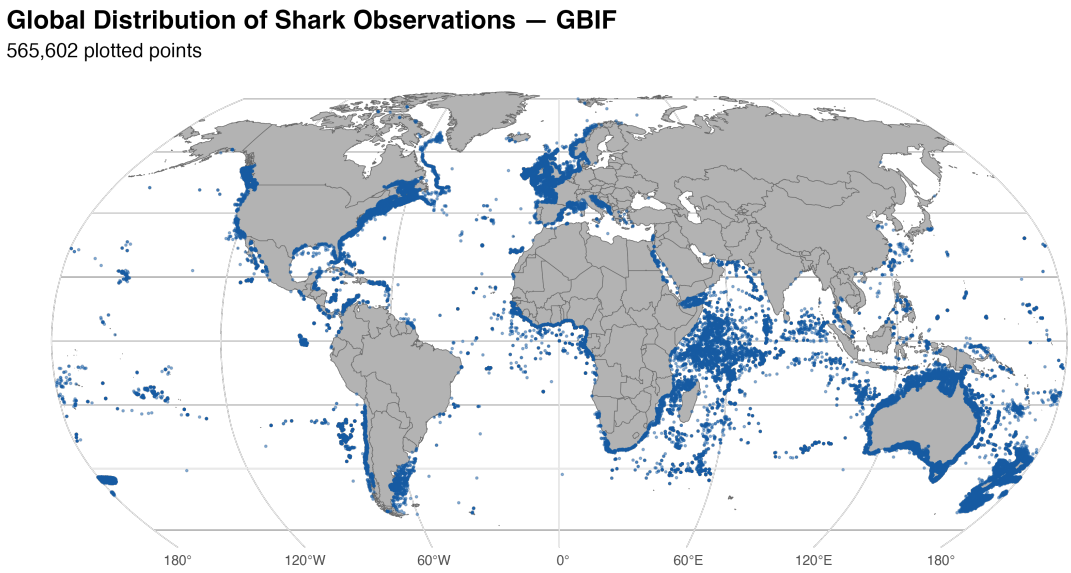


Figure 4.7: Global distribution of GBIF shark observations.

### 4.3.5 SPUE trends

When grouped spatially by the two case-study regions, platform SPUE predictions often showed similar trends, suggesting that they both captured their respective user engagement. Flickr trends should be interpreted cautiously given uncertainties in automatic annotation, whereas iNat indices provide a stronger ecological baseline.

Modeled SPUE trends show annual mean values (points) with 95% confidence intervals (error bars) and fitted GLM trajectories (solid lines with shaded 95% confidence ribbons) (Figures 4.8, 4.9, 4.10, and 4.10). An example of observed data and model diagnostics are presented in Appendix G to exhibit a typical modeled species' goodness-of-fit.

### Flickr

To reveal taxonomic confidence of predicted SPUE trends, we appended the SD's species-specific classification accuracy for each modeled species (Figures 4.8 and 4.9).

**Bahamas.** Species-specific SPUE trends from Flickr indicate heterogeneous population trajectories among seven focal shark species across the Bahamian archipelago (Figure 4.8). Reef-associated taxa such as the nurse shark (*Ginglymostoma cirratum*) and whitetip reef shark (*Triaenodon obesus*) exhibited modest annual declines of –4.9% (95% CI: –9.4 to –0.2%) and –7.3% (95% CI: –14.8 to 0.8%), respectively. The tiger shark (*Galeocerdo cuvier*) showed a stronger negative trend of –9.2% per year (95% CI: –16.1 to –1.8%). In contrast, coastal species including the Caribbean reef shark (*Carcharhinus perezii*) and bull shark (*C. leucas*) displayed weak positive trajectories of 6.4% (95% CI: –4.1 to 18.0%) and 6.8% (95% CI: –4.5 to 19.4%) per year, respectively. The blacktip reef shark (*C. melanopterus*) exhibited the clearest increase, rising by 13.7% annually (95% CI: 3.4 to 25.1%). Collectively, these trends suggest stable to increasing encounters for nearshore *Carcharhinus* species and modest declines among larger, wide-ranging taxa, underscoring spatial and ecological contrasts in shark occurrence patterns throughout the Bahamas.

**Hawaii.** Species-specific SPUE trends from Flickr for Hawaii indicate lower reporting intensity and greater uncertainty than in the Bahamas, reflecting comparatively fewer shark observations across the archipelago. Reef-associated taxa displayed mixed or declining trajectories. The grey reef shark (*Carcharhinus amblyrhynchos*) and Galapagos shark (*C. galapagensis*) declined markedly by –19.5% (95% CI: –27.8 to –10.3%) and –15.5% (95% CI: –29.3 to 0.9%) per year, respectively. The whitetip reef shark exhibited a modest, uncertain increase of 2.4% per year (95% CI: –1.9 to 6.8%). Among pelagic species, the tiger

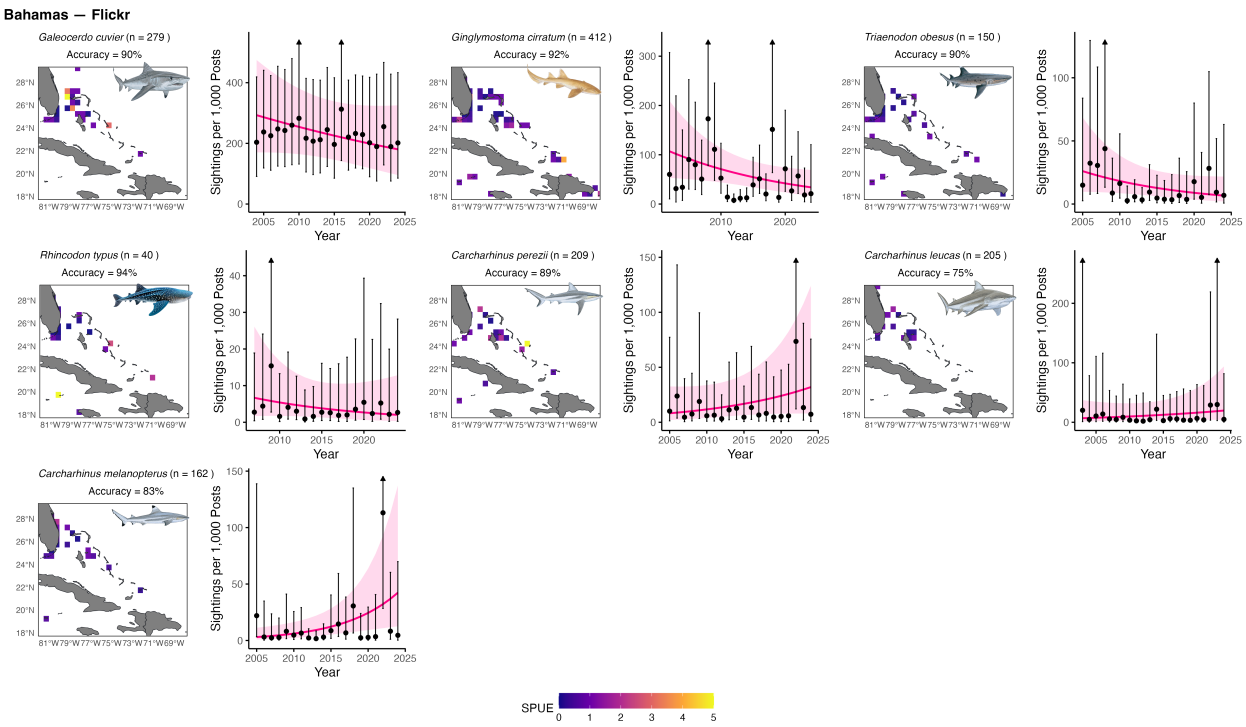


Figure 4.8: Flickr SPUE trends exhibited for seven shark species in the Bahamas. The points and error bars represent interannual predictions and variability. Upward arrows represent extended arrow bars, while the pink-colored line and ribbon represent the platform-specific colored long-term predictions and uncertainty within a 95% confidence interval.

shark and whale shark both showed shallow but uncertain declines of –5.1% (95% CI: –15.1 to 6.0%) and –7.3% (95% CI: –20.3 to 7.8%), respectively. The scalloped hammerhead (*Sphyrna lewini*) exhibited a moderate annual decline of –6.2% (95% CI: –11.7 to –0.2%). Collectively, these results point to generally decreasing or variable encounter frequencies for both reef-associated and pelagic sharks in Hawaii, consistent with reduced observation effort and greater temporal variability relative to the Bahamas.

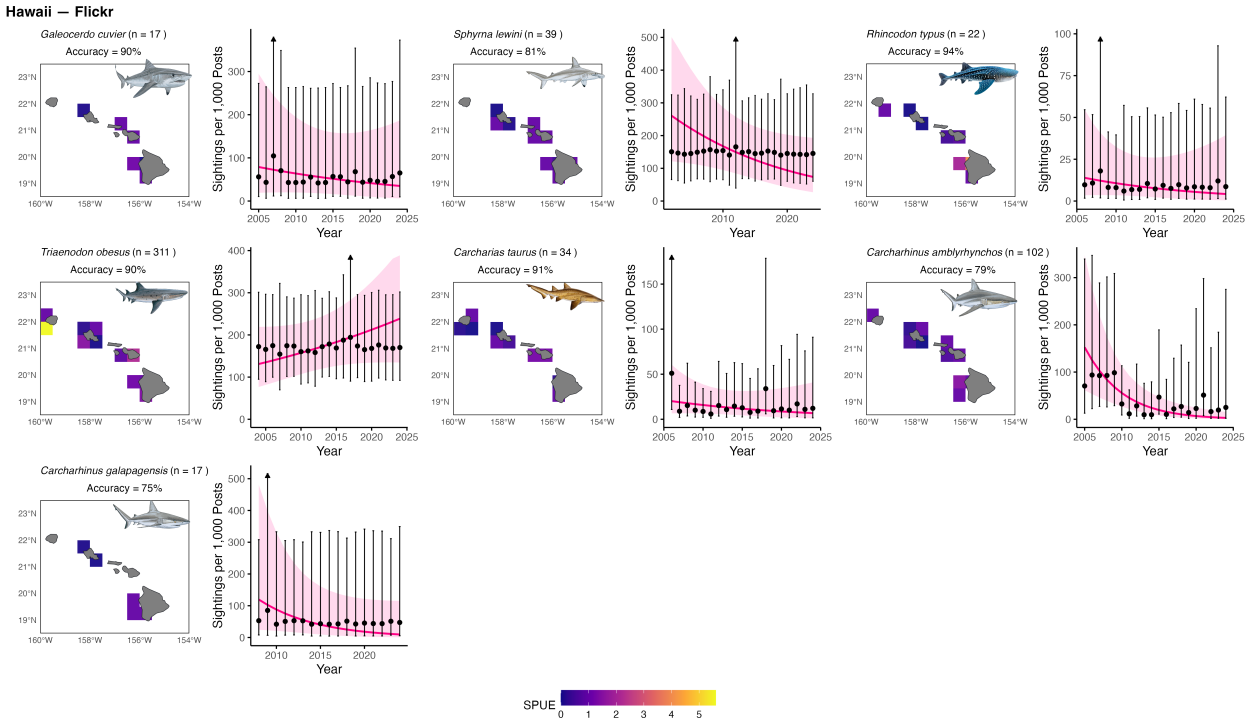


Figure 4.9: Flickr SPUE trends for seven shark species in the MHIs. The points and error bars represent interannual predictions and variability. Upward arrows represent extended arrow bars, while the pink-colored line and ribbon represent the platform-specific colored long-term predictions and uncertainty within a 95% confidence interval.

### iNaturalist

**Bahamas.** Species-specific SPUE trends from iNat show strong and consistent increases across all nine focal shark taxa in the Bahamas. Reef-associated species such as the nurse

shark and lemon shark (*Negaprion brevirostris*) increased by 13.9% (95% CI: 10.2 to 17.8%) and 35.7% (95% CI: 27.2 to 44.9%) per year, respectively, while the bull shark and blacktip shark (*C. limbatus*) exhibited similar or stronger increases of 36.2% (95% CI: 24.9 to 48.6%) and 42.7% (95% CI: 30.4 to 56.2%) per year. The tiger shark and great hammerhead shark (*Sphyrna mokarran*) also rose sharply by 20.7% (95% CI: 8.8 to 33.8%) and 18.8% (95% CI: 7.0 to 32.0%) annually. Smaller-bodied coastal species, including the bonnethead shark (*Sphyrna tiburo*) and Atlantic sharpnose shark (*Rhizoprionodon terraenovae*), increased by 12.4% (95% CI: 6.6 to 18.4%) and 33.5% (95% CI: 17.9 to 51.2%) per year, respectively. The silky shark (*C. falciformis*) exhibited the strongest rise, increasing by 56.6% annually (95% CI: 16.5 to 110.4%). Collectively, these results indicate broad and substantial upward trends in reported shark encounters across both coastal and pelagic taxa, reflecting rising observation rates and potentially increasing shark presence throughout the Bahamian archipelago.

**Hawaii.** Species-specific SPUE trends from iNat in Hawaii reveal predominantly stable to declining trajectories across most taxa. Reef-associated species exhibited the strongest negative trends: the scalloped hammerhead shark declined by –10.9% per year (95% CI: –19.5 to –1.5%), and the grey reef shark decreased by –14.0% annually (95% CI: –23.3 to –3.5%). The blacktip reef shark showed a weak, uncertain decline of –4.4% (95% CI: –12.3 to 4.3%), while the whitetip reef shark increased modestly by 6.5% (95% CI: 2.9 to 10.1%) per year. Among wide-ranging or seasonally coastal taxa, the tiger shark showed a near-stable trajectory (–0.5% per year: 95% CI: –9.0 to 8.8%), and the whale shark declined slightly by –6.4% (95% CI: –16.1 to 4.4%). The Galapagos shark exhibited a weak positive trend of 7.7% (95% CI: –0.5 to 16.6%), while the blacktip shark increased by 10.5% annually (95% CI: –6.3 to 30.3%). Overall, these trends indicate generally lower or variable encounter rates for reef-associated sharks near the MHIs, contrasting with occasional increases among

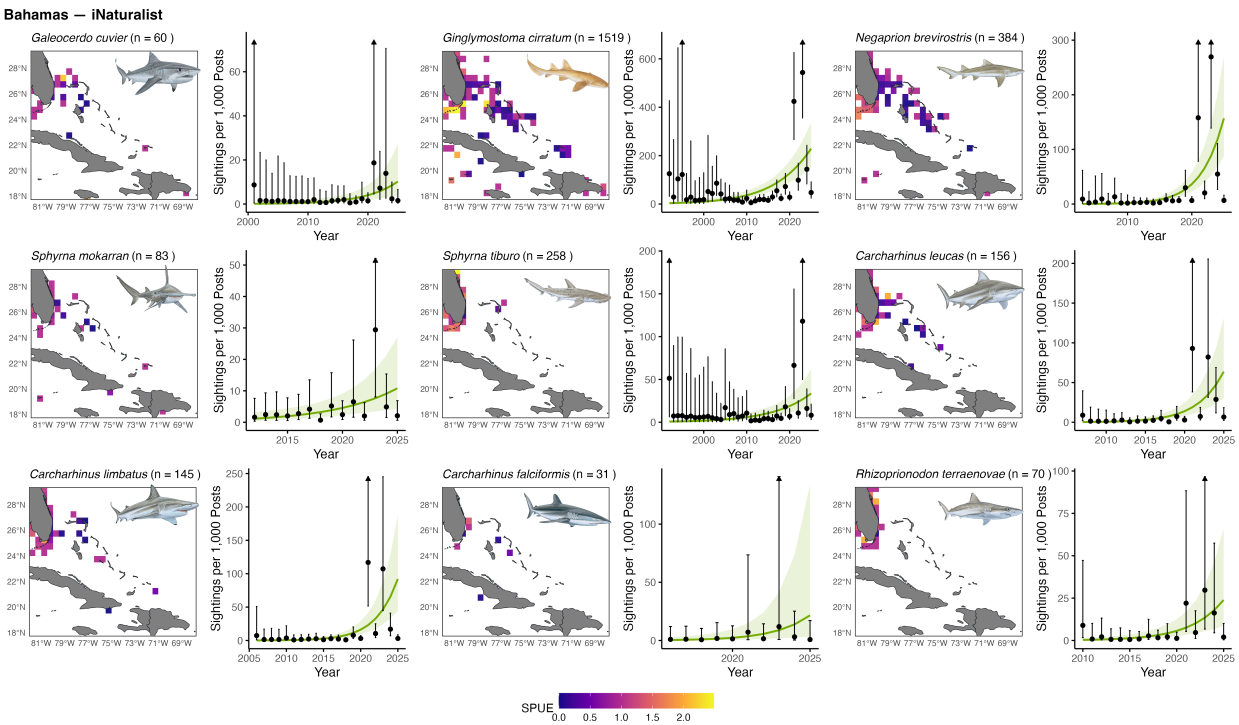


Figure 4.10: iNat SPUE predictions for nine species in the Bahamas. The points and error bars represent interannual predictions and variability. Upward arrows represent extended arrow bars, while the green-colored line and ribbon represent the platform-specific colored long-term predictions and uncertainty within a 95% confidence interval.

more mobile coastal and pelagic species.

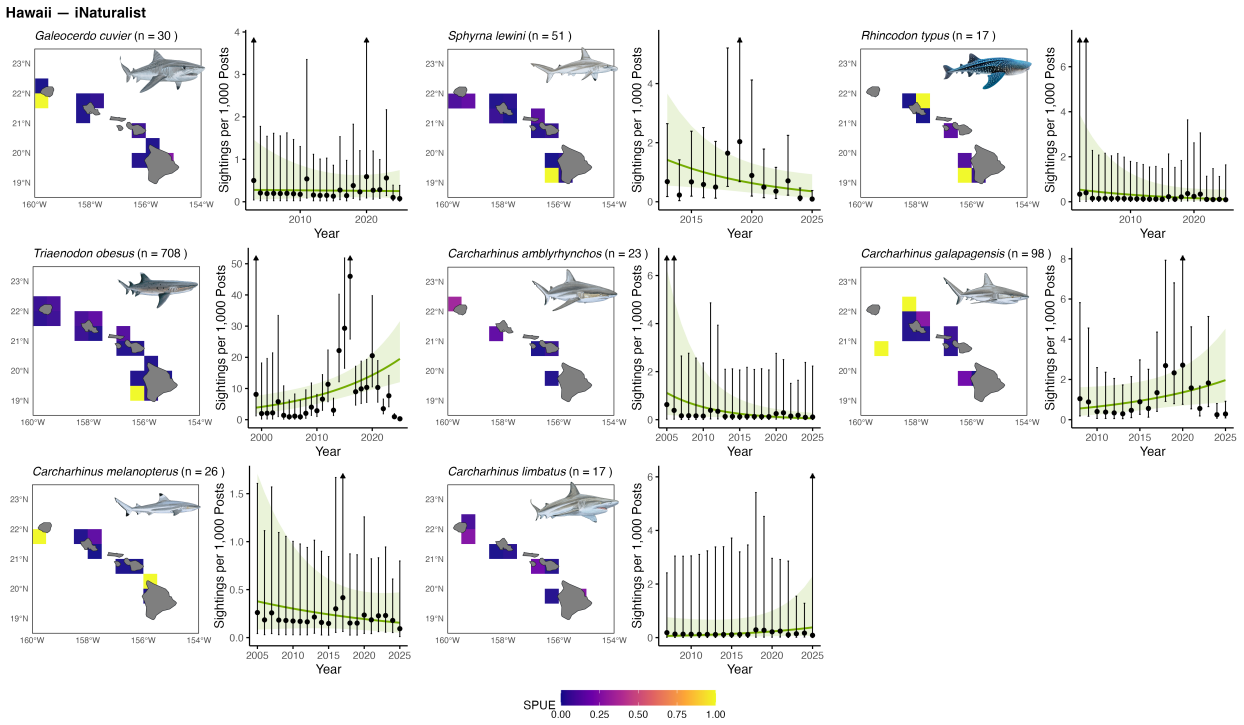


Figure 4.11: iNat SPUE predictions for eight species in the MHIs. The points and error bars represent interannual predictions and variability. Upward arrows represent extended arrow bars, while the green-colored line and ribbon represent the platform-specific colored long-term predictions and uncertainty within a 95% confidence interval.

Across both platforms, SPUE trends reveal clear regional contrasts in shark occurrence patterns. In the Bahamas, Flickr and iNat data consistently indicate strong increases across most species and ecological groups, reflecting recurrent observations of reef-associated and coastal taxa. In contrast, Hawaii shows predominantly stable or declining trajectories, particularly among reef-associated species, with only a few wide-ranging taxa exhibiting weak or uncertain increases. Together, these findings demonstrate that SN SPUE indices can detect broad-scale differences in shark population trajectories and observation effort between regions.

### 4.3.6 Comparative analysis

We sourced independent abundance assessments that report the following indices: research longline CPUE, BRUV MaxN, and standardized visual encounter surveys (Table 4.1). The Bahamas literature is dominated by coastal longline programs [65, 81, 104, 124], opportunistic and acoustic studies [148], and BRUV monitoring across protected and non-protected sites [13, 24, 51]. Hawaii sources emphasize longline catch data [39], extensive BRUV and visual surveys [5, 6, 7, 8, 37], and tourism or telemetry studies documenting tiger and reef shark encounters [51, 105, 106, 141].

Overall, directional agreement between SPUE and published indices was strongest for (i) coastal Bahamian species with sustained monitoring, and (ii) the well-documented decline of reef-associated sharks in Hawaii relative to the Northwestern Hawaiian Islands. Flickr trends disagreed more strongly than iNat trends, and were often attributable to differences in sampling scope (e.g., nearshore SPUE versus offshore longline CPUE) and uncertainty in automatic taxonomic classification for Flickr observations. Several taxa (particularly hammerheads) showed mixed or uncertain patterns due to limited temporal coverage and low observation encounters.

**Bahamas comparisons.** iNat SPUE trends for the Bahamas show widespread increases among coastal and reef-associated species such as tiger, lemon, nurse, and hammerhead sharks (great and bonnethead). These directions align with longline and BRUV programs reporting stable or increasing abundance during overlapping years [24, 65, 81, 104, 120] (Table 4.1). In particular, Tiger Beach has become a predictable aggregation site for female tiger sharks, where protection under the national shark sanctuary (since 2011) and regulated ecotourism may reinforce continued high encounter rates [13, 148].

**Hawaii comparisons.** In Hawaii, SPUE trends largely mirror independent observations showing reduced reef-shark abundance around populated MHI reefs compared to the Northwestern islands [6, 8, 51, 141]. Declines in whitetip, blacktip, grey reef, Galapagos, and sandbar sharks are consistent across sources, whereas tiger sharks exhibit stable to slightly increasing trends, matching telemetry and encounter studies that highlight frequent coastal residency near human activity [106] (Table 4.1). Flickr and iNat report declining scalloped hammerhead populations, congruent with longline surveys [39].

Together, these comparisons show that SN SPUE indices broadly reproduce known regional population patterns—detecting coastal shark recovery and protection effects in the Bahamas, and reef-shark depletion near developed islands in Hawaii—while remaining sensitive to data gaps and observational biases.

| Region | Species / Group | Flickr SPUE | iNat SPUE | Reference Trend | Agreement (iNat & Reference) | Method Type |
|---|---|---|---|---|---|---|
| Bahamas | *Galeocerdo cuvier* (Tiger) | ↓ | ↑ | ↑ | Agree | Longline (CPUE) |
| Bahamas | *Negaprion brevirostris* (Lemon) | ? | ↑ | ↑ | Agree | Longline (CPUE) |
| Bahamas | *Ginglymostoma cirratum* (Nurse) | ↓ | ↑ | ≈ / ↑ | Mixed | BRUVs (MaxN) |
| Bahamas | *Sphyrna mokarran* (Great hammerhead) | ? | ↑ | ↑ | Agree | Mixed (CPUE + BRUVs) |
| Bahamas | *Sphyrna tiburo* (Bonnethead) | ? | ↑ | ↑ | Agree | Longline (CPUE) |
| Bahamas | Reef sharks (Whitetip, Blacktip, Grey, Caribbean) | ≈ | ↑ | ↑ / ≈ | Mixed | BRUVs / Visual |
| Hawaii | Reef sharks (Whitetip, Blacktip, Grey, Galapagos, Sandbar) | mixed | ↓ | ↓ | Agree | BRUVs / Visual |
| Hawaii | *Galeocerdo cuvier* (Tiger) | ≈ | ↑ | ≈ / ↑ | Mixed | Visual / Telemetry |
| Hawaii | *Sphyrna lewini* (Scalloped hammerhead) | ↓ | ↓ | ↓ | Agree | Longline (CPUE) |
| Hawaii | *Carcharhinus amblyrhynchos* (Grey reef) | ↓ | ↓ | ↓ | Agree | BRUVs (MaxN) |

Table 4.1: Comparison of Flickr and iNat SPUE trends with independent reference assessments from longline, BRUV, and visual encounter studies. Agreement categories denote directional consistency between SPUE and published indices: **Agree** = consistent trend direction: **Mixed** = partial or conflicting evidence: **Unknown** = insufficient data. When a trend direction is marked with ≈, it denotes a weak or stable trend. If combined with another directional symbol, it indicates that a secondary reference study and/or method (e.g., CPUE, BRUV, or SPUE) provided additional evidence for the same regional population.

## 4.4 Discussion

Our results show that large, heterogeneous digital observation streams can be converted into indices of relative abundance that reveal patterns in shark populations across regions. By pairing platform records with effort indicators, we modeled counts with a negative binomial distribution and obtained SPUE time series that align in direction with independent assessments. In the Bahamas, iNat indicate multispecies increases among coastal and reef taxa, consistent with protection and predictable aggregation sites reported elsewhere [81, 104, 148]. Around the MHIs, reef-associated species tend to be stable or declining (shown by iNat and Flickr) while some wide-ranging taxa appear flatter, echoing contrasts between populated and remote islands documented by BRUVs, visual surveys, and telemetry [6, 8, 106]. These agreements are not uniform and should not be read as replacements for CPUE or BRUVs: rather, they indicate that filtered posts, once standardized by effort, can complement conventional monitoring methods at low cost.

Platform differences explain much of the remaining divergence. iNat offers community-vetted taxonomy and consistent spatiotemporal fields, which produced the most stable SPUE estimates in our case studies. Flickr's deep archive and metadata are valuable, but automatic species labels and more touristic behavior introduce annotation and participation bias that widen uncertainty. Our expert rechecks improve reliability, yet this is manually tedious without stronger citizen science engagement. IG contains unmatched volumes of recent shark encounters but limited programmatic access prevents reproducible effort standardization, constraining current utility for formal indices despite clear scientific potential. While these raw observations are currently challenging to standardize, they are still useful for training the SD, boosting its taxonomic range and classification performance. With further standardization of GBIF's ingested systematic surveys, incorporating presence/absence indices will further supplement the value of this workflow. Across all platforms, visibility bias to-

ward charismatic species and popular sites, uneven participation across seasons and regions, and mismatches with reference series (for example, nearshore SPUE versus offshore longline CPUE) encompass the limitations. Within these limits, SPUE should be viewed as an index of encounter rate that primarily reflects changes in direction, timing, and spatial differences in relative abundance.

The Bahamas patterns illustrate how management and human activity can shape detectability in ways that are still ecologically meaningful. The sanctuary designation and long-standing limits on commercial longlining align with increasing SPUE for multiple taxa, including charismatic and coastal species frequently encountered by observers. In Hawaii, the consistent declines in reef-associated sharks near the MHIs mirror independent evidence for lower predator densities around populated islands relative to the Northwestern Hawaiian Islands, with wide-ranging species showing flatter or uncertain trajectories. These contrasts suggest that citizen science observations, when modeled with effort offsets and seasonality, can recover real differences in relative abundance even when absolute abundance is unknown.

A tangible example is the bonnethead shark in the Northwest Atlantic region. Our iNat SPUE indicates a clear increase in the eastern Florida–Bahamas neighborhood during the observation window (about +12.4% per year: 95% CI 6.6 to 18.4%). This local rise is consistent with parts of the western Atlantic where recent assessments report strong positive regional signals, while longer series for the broader South Atlantic and Gulf show overall declines [120]. SPUE indices provide fine spatiotemporal resolution that helps identify where relative abundance is increasing and where it is not. This additional layer of information can complement local and regional stock assessments by guiding validation surveys and clarifying heterogeneous trends across neighboring stocks, without assuming uniform population dynamics at broader spatial scales.

Methodologically, our approach aims to make bias explicit rather than hidden. Treat-

ing year as either continuous or independent isolates long-term patterns from year-to-year variability, while including seasonal terms captures cyclic detectability common to both animal activity and observer behavior. Incorporating years with zero observations retains the information that effort-standardized encounters were rare and ensures that uncertainty is represented transparently. Where reference time series target habitats not well sampled by citizen observers, partial disagreement is expected: aligning spatial and seasonal bins to improve ecological comparability remains the logical next step.

More broadly, the framework presented here serves as a blueprint for integrating heterogeneous digital observations into ecological monitoring. Its value lies in consolidating well-defined occurrence records from sources that have been underutilized in wildlife science and demonstrating a statistical pathway for transforming them into indicators of relative abundance. The modeling structure we applied is a starting point that can be refined to better account for the biases and uncertainties already identified in this study. Beyond sharks, this workflow can be extended to other marine and terrestrial animal groups captured through citizen and research-driven digital reporting, offering a scalable and transparent foundation for future biodiversity monitoring.

In practical terms, SPUE is immediately useful for agencies and programs confronting data gaps. Directional signals can identify emerging increases or declines ahead of formal surveys, guide where limited field effort can have the greatest value, and provide context for interpreting CPUE or BRUVs when those indices shift or lack scale. In data-deficient settings, particularly for threatened species with sparse or inconsistent monitoring, standardized encounter-rate trends offer a scalable tool for detecting emerging population changes and guiding where traditional surveys are most urgently needed. By flagging potential declines or recoveries in near real time, these indices can help prioritize limited monitoring resources and trigger management responses before data gaps become critical. The frame-

work demonstrated here establishes a foundation for integrating digital observation streams with conventional programs, ensuring that citizen and open-data signals are not isolated but instead contribute meaningfully to management decision-making. With transparent effort standardization, explicit uncertainty reporting, and continued validation, these approaches can establish social media as digital monitoring networks, informing stock assessments, and strengthening the evidence base for conservation policy.

# Chapter 5

# Detecting Mediterranean White Sharks and Broader Elasmobranch Biodiversity with Environmental DNA

Published as J. F. Jenrette, J. Jenrette, N. Kobun Truelove, S. Moro, N. Dunn, T. Chapple, A. Gallagher, C. Gambardella, R. Schallert, B. Shea, D. Curnick, B. Block, and F. Ferretti. Detecting Mediterranean White Sharks with Environmental DNA. *Oceanography*, 3 2023. URL https://doi.org/10.5670/oceanog.2023.s1.28

The material presented here is adapted from the published article with substantial updates, supplementary materials, and methodological clarifications.

# Abstract

Elasmobranchs across the Mediterranean Sea have suffered some of the steepest population declines of any marine vertebrates, driven by centuries of overfishing, bycatch, and habitat degradation. Many species are now regionally extinct or persist at densities too low to be detected by conventional surveys. Among them, the white shark (*Carcharodon carcharias*) is the most critically endangered and least observed, leaving its status and distribution unresolved for decades. To overcome these limitations, we implemented an integrative monitoring framework from 2021–2024 that combined Environmental DNA (eDNA) analyses, oceanographic particle modeling, and citizen science to detect and track white shark presence across the Mediterranean Sea without the need to directly observe them. A total of 204 seawater samples were collected from 11 regions spanning the Sicilian Channel, Tunisian Plateau, and adjacent basins. Species-specific assays detected white shark eDNA at five sites, confirming population persistence. Particle tracking simulations indicated that detections corresponded to the animal shedding eDNA within roughly 48 hours prior and 20–25 km of sampling, enabling near-real-time tracking of individual movement. Laboratory controls, field blanks, and tissue-derived reference samples ensured stringent quality assurance and minimized false detections. Broader metabarcoding of 48 samples revealed 12 elasmobranch taxa, including all three Mediterranean Lamnids—white shark, shortfin mako (*Isurus oxyrinchus*), and porbeagle (*Lamna nasus*). Citizen scientists extended spatial coverage across the Ligurian, Tyrrhenian, and Adriatic Seas, demonstrating the scalability of this approach. Together, these complementary methods establish a reproducible molecular and citizen-participation blueprint for detecting and monitoring critically endangered elasmobranchs.

## 5.1 Introduction

Sharks are vital ecological regulators that maintain the structure and stability of marine ecosystems, yet throughout the Mediterranean Sea they face an unprecedented conservation crisis [112, 118]. Once common across coastal and pelagic habitats, Mediterranean elasmobranchs have undergone steep population declines due to centuries of overfishing, bycatch, habitat degradation, and climate-driven shifts in distribution [112]. Today, approximately 65% of shark and ray species in the region are considered threatened with extinction [45]. This is a dramatic rise from 29% in 1980. Additionally, 27.5% of shark species remain data-deficient [48]. This highlights the dual challenge of both overexploitation and inadequate monitoring, as inconsistent and absent reporting hinders robust population assessments [112]. Within this broader context, the white shark (*Carcharodon carcharias*) stands as the most extreme example of decline and data deficiency [54]. Listed as Critically Endangered by the International Union for Conservation of Nature (IUCN) [42], the Mediterranean population has been reduced by an estimated 52–96% relative to historical levels [54, 111]. Conventional monitoring with catch records, diver surveys, and electronic tagging is typical for informing conservation actions and reversing population declines. However, Mediterranean white sharks are now reduced to the brink of extinction, making their detection and monitoring extremely difficult with such methods [54]. Innovative approaches capable of detecting sharks indirectly and guiding field effort toward likely encounter zones are therefore essential.

Environmental DNA (eDNA) analyses have emerged as one of the most transformative tools in modern biodiversity monitoring. Rooted in molecular forensics, it relies on detecting trace genetic material—cells, mucus, or metabolic waste—that organisms shed into their surroundings. By capturing and amplifying this environmental genetic signal, researchers can infer the presence of species without the need to visually observe, capture, or disturb them. The approach has revolutionized wildlife surveys across ecosystems [16], from de-

tecting amphibians in freshwater ponds to tracking cryptic terrestrial mammals and large marine vertebrates, including sharks, whales, and sea turtles [19]. Its power lies in sensitivity and scalability: a single liter of seawater can contain genetic traces of numerous species, providing a snapshot of local biodiversity that traditional surveys would miss.

For the Mediterranean Sea, where many elasmobranchs are critically depleted and rarely encountered, eDNA represents a crucial advancement [85]. It offers a non-invasive, cost-effective, and replicable method to reveal species occurrence across broad spatial and temporal scales—precisely where conventional surveys fail. In regions where direct observation is improbable, such as the deep or pelagic habitats occupied by the last remaining white sharks, eDNA provides a new avenue to locate recent presence [3], prioritize search areas, and guide adaptive fieldwork. This study was initially driven to apply these methods exclusively to the white shark [80], and then subsequent studies aimed to characterize broader elasmobranch biodiversity. With molecular assays designed with species-specificity and multi-species targets, detection can extend beyond observation to inform strategic conservation by highlighting the last strongholds of this threatened group of marine animals in the Mediterranean Sea.

Species-specific assays provide the necessary precision to detect target taxa with high confidence, an especially critical feature when distinguishing between closely related species [36]. In the Mediterranean Sea, this distinction is vital: the critically endangered white shark and its more common relative, the shortfin mako (*Isurus oxyrinchus*), share much of the same habitat and resemble one another, making unverified detections prone to misclassification [161]. By tailoring molecular markers to the white shark's unique mitochondrial DNA sequence, researchers can reduce False Positive (FP) detections and identify rare species, with confidence, in a mixed pelagic environment.

A second advancement in eDNA analyses lies in linking field-based detection with labora-

tory confirmation and coupling molecular data with oceanographic modeling [38]. Because vessel operations are costly and encounter rates are low, the ability to detect genetic signals directly at sea offers a strategic advantage. Rapid, qualitative screening of eDNA can provide near-real-time awareness of species presence, guiding where and when to intensify sampling or deploy other tools such as cameras or electronic tagging gear [54]. Beyond this, the development of oceanic particle tracking simulations allows researchers to reconstruct the likely origin and trajectory of detected genetic material. By modeling how eDNA molecules drift and degrade in seawater, these hindcast analyses offer a spatiotemporal snapshot of where individuals were likely roaming prior to detection. Open-source frameworks such as `OpenDrift` [38] enable these simulations to incorporate eDNA-specific parameters including degradation rate, buoyancy, and current velocity [3, 138] to forecast areas of recent species presence. When paired with laboratory confirmation, these tools transform eDNA from a purely retrospective survey method into a dynamic monitoring framework capable of informing adaptive field strategies and guiding future search effort [4, 46].

Scaling such detection across time and space requires broad participation. Citizen science networks can play a pivotal role in expanding eDNA monitoring capacity beyond the limited footprint of research expeditions. By distributing standardized sampling kits to sailors, divers, and ocean-goers, scientists can significantly increase the chance for detecting rare and threatened species [110]. Further, established and consistent citizen engagement can generate long-term biodiversity patterns [67, 149].

Metabarcoding and multiplexing approaches complement these targeted assays by capturing the wider elasmobranch community [163]. Although less precise at the species level, metabarcoding reveals the ecological context in which critically endangered species persist, identifying co-occurring taxa and potential hotspots of shared vulnerability [109, 110].

While these approaches collectively expand the frontier of marine molecular monitoring,

they are not without limitations. The sensitivity that makes eDNA powerful also renders it susceptible to contamination and sampling error [109]. FP detections may arise from trace contamination, detection of non-target DNA, or confounding degraded genetic material [161]. Similarly, False Negative (FN) detections can occur when environmental conditions accelerate DNA degradation [34] or when local concentrations fall below detection thresholds [94, 161], which is common when sampling water in pelagic environments. Field protocols must therefore emphasize rigorous sterilization and replication, including the use of field blanks, negative controls, and standardized sampling volumes to ensure reliability [36]. Citizen science sampling introduces an additional layer of uncertainty, as inconsistent handling or storage can compromise sample integrity [110]. Tailored bioinformatic pipelines incorporating positive controls and quantification of uncertainty are essential for both species-specific and metabarcoding assays [61, 133].

In this study, we address these limitations by establishing a reproducible framework for eDNA monitoring in data-poor marine systems. By integrating field and laboratory approaches, combining species-specific and multi-species assays, and expanding sampling through coordinated citizen participation, we provide a comprehensive roadmap for detecting and tracking critically endangered elasmobranchs in the Mediterranean Sea. Over four years (2021–2024), we collected more than 200 environmental samples, including contributions from trained citizen scientists, across 11 distinct Mediterranean regions. Our workflow incorporated quality controls such as tissue-derived positive controls, and both field and laboratory blanks, to validate detections and quantify uncertainty. Preliminary analyses revealed the presence of 12 elasmobranch species, including all three critically endangered Lamnid sharks known from the Mediterranean Sea. Together, these efforts demonstrate eDNA as an evolving supplementary monitoring method for detecting the region's most imperiled marine predators.

## 5.2 Methods

### 5.2.1 Field Sampling

Field sampling was conducted between 2021 and 2024 across Mediterranean regions identified as either historical white shark hotspots or opportunistic sampling areas (Figure 5.7). Sampling locations were selected using historical sighting records [52, 111, 118], with additional stations chosen opportunistically to maximize use of ship time and increase geographic coverage. This dual strategy balanced systematic targeting with adaptive sampling to improve spatial representation across regions where white sharks are rarely observed.

Sampling was performed in transects while research vessels were in transit to stations for Baited Remote Underwater Video (BRUV) deployments, drone-based searches, and chumming operations. Seawater samples were collected in duplicate (and occasionally triplicate) from 0–100 m depth using a 5 L Niskin bottle, with 2–5 L of seawater filtered per sample (Figure 5.1A).

Filtration was carried out using one of three standardized methods: (i) vacuum manifold (Figure 5.1B) or hand-pump filtration through 0.45 µm Polyvinylidene Fluoride (PVDF) filter paper, (ii) Syringe filtration through Sterivex filter cartridges [88], or (iii) citizen-science kits (Figure 5.2d-f) containing a siphon pump and self-preserving filters [157]. Each method was designed to capture eDNA fragments suspended in seawater while accommodating varying field conditions and available equipment. After filtration, filters were either preserved in DNA/RNA Shield (Zymo Research) or RNAlater (Thermo Fisher Scientific) reagent for laboratory processing, or used immediately for rapid onboard eDNA analysis targeting white shark detection.

All filtration equipment, Niskin bottles, and work surfaces were sterilized between sam-

ples using 5% bleach and 70% ethanol. Field blanks (one per sampling day when conditions allowed) were included throughout sampling to monitor contamination. Duplicate and triplicate samples were taken routinely to evaluate repeatability and detection consistency. Preserved samples were stored at –20 °C prior to extraction and sequencing at the Virginia Tech Genomics Sequencing Center (VT-GSC).

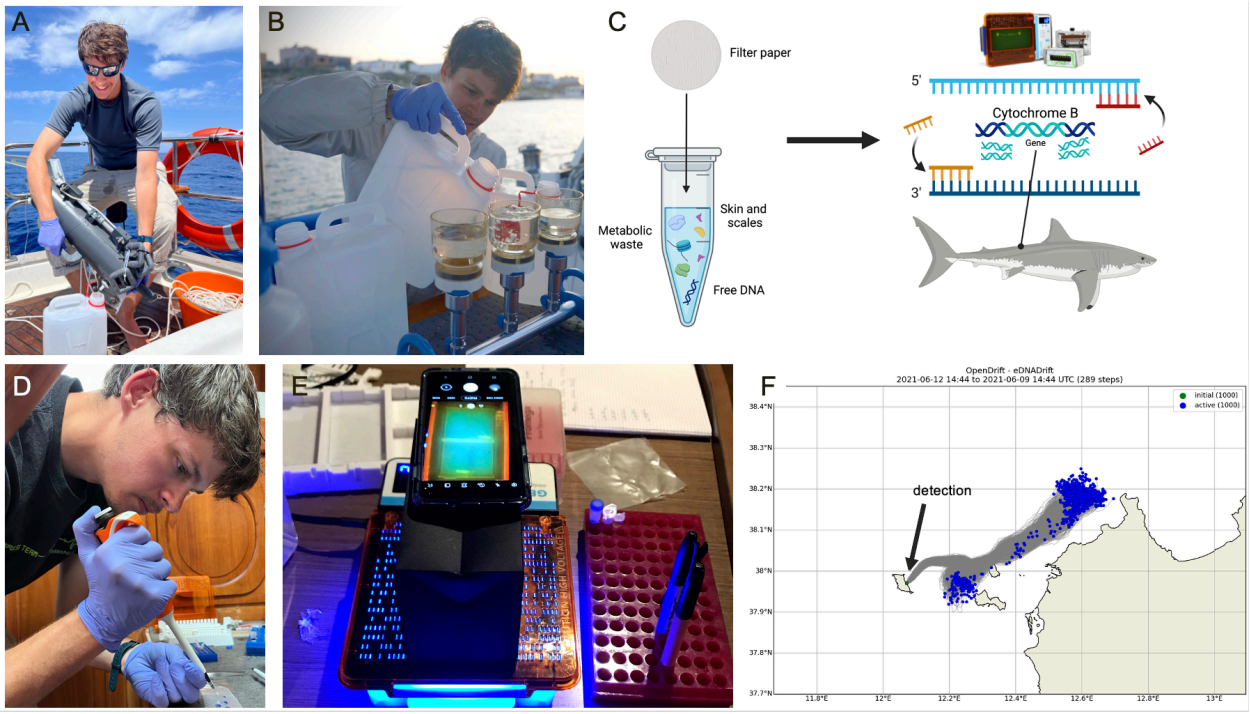


Figure 5.1: Workflow of the eDNA detection pipeline. (A) Collection of 2–5 L of seawater from 0–100 m depth. (B) Filtration of three water samples simultaneously using a vacuum manifold apparatus. (C) Cell lysis and DNA extraction followed by Polymerase Chain Reaction (PCR) amplification of the white-shark-specific mitochondrial gene. (D) Preparation of amplified samples for visualization via gel electrophoresis. (E) Validation of electrophoresis results through sequencing. (F) Particle dispersal hindcasting predicting the origin of eDNA shedding from the latest positive white shark detection: the green initial particle represents the detection site, and blue active particles indicate the backward trajectory simulated under current velocity and water temperature conditions.

### 5.2.2 Citizen Science

Kits were distributed to volunteer vessels from 2023-2024. Each kit included a manual siphon pump, five self-preserving 0.45 µm filter cartridges designed in Thomas et al. [157], gloves, and a data sheet for recording date, location, depth, and sampler name (Figure 5.2). Seven kits were deployed in the Ligurian, Tyrrhenian, and Adriatic Seas by five different participating citizens. More kits are being currently deployed in Croatia and the Aegean Sea. Used filters were mailed to VT-GSC for analysis under the same molecular pipeline.

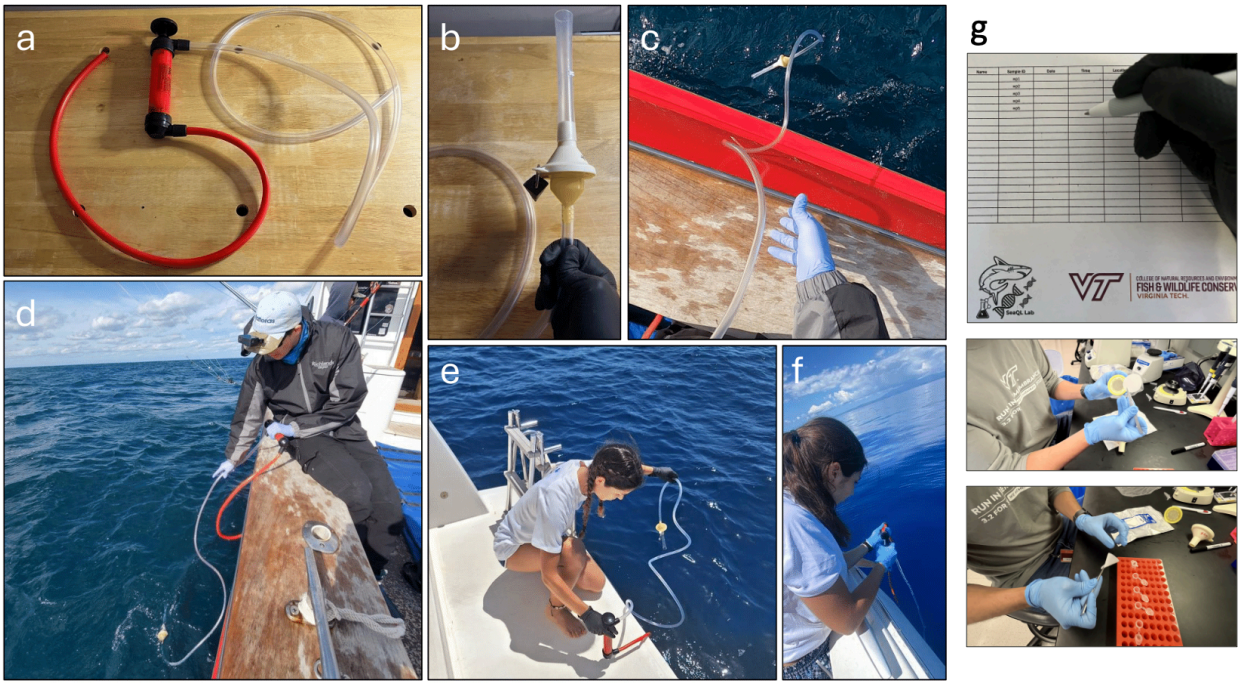


Figure 5.2: Citizen science eDNA sampling kit and deployment workflow. Each kit enables users to filter 2 L of surface seawater using (a) a manual siphon pump with inlet and outlet tubing, and (b) self-preserving filter units with 0.45 µm pores and latex gloves. Panels (c) and (d) show the sampling procedure as instructed in the user manual: the filter is firmly attached to the siphon inlet and lowered just below the sea surface. In (e), a participating volunteer collects a sample off of northern Sardinia: in (f), another volunteer pumps seawater through the filter. Following filtration, users are instructed to return the filter and data sheet to its original package and ship it back to the VT-GSC for laboratory processing in (g).

### 5.2.3 DNA Extraction and Amplification

eDNA was extracted from filters either onboard or in the laboratory depending upon time and logistical constraints. Onboard analyses used the RNAGEM V kit (MicroGEM), a rapid extraction method that lyses cellular material and produces DNA-ready solutions within 15 min at 75 °C, yielding approximately 1 mL of extract per sample [128] (Figure 5.1C). This rapid workflow enabled near-real-time genetic screening while the vessel was still at sea. For a more comprehensive extraction in controlled laboratory conditions, the DNeasy Blood & Tissue Kit (Qiagen) [121] was employed, which provides higher yields through a one-hour lysis and purification process suitable for downstream sequencing.

For species-specific detection of the white shark, amplification targeted a 151 bp fragment of the mitochondrial Cytochrome B (CYTB) gene using primers developed by Lafferty et al. [94] (*forward*: 5- CGTCACCCCTCCACACATTA -3 : *reverse*: 5- GGTGCTGC-TACGTTGTTTGG -3 ). These primers were selected for their specificity to white shark DNA and low cross-detection with the shortfin mako, which shares approximately 89% sequence similarity at the CYTB gene [80]. Other white shark mitochondrial regions (e.g., *COI*, *ND2*) can serve as alternative targets to further minimize false positives or negatives when assay refinement is necessary. PCR assays were prepared with Platinum SuperFi II master mix (Invitrogen) in 25 µL reactions containing 0.5 µM of each primer and 2–10 µL of DNA template. Thermal cycling consisted of an initial denaturation at 94 °C for 3 min followed by 40 cycles of 94 °C for 30 s, 52 °C for 30 s, and 72 °C for 30 s, with a final extension at 72 °C for 1 min. Amplicons were visualized on a miniPCR GELATO electrophoresis system onboard (Figure 5.1D-E) and re-validated at the VT-GSC using an Agilent TapeStation.

For broader elasmobranch biodiversity assessment, we applied a metabarcoding approach using the MiFish-E primer set [46, 109]. This universal 12S rRNA primer pair (*forward*:

5- GTCGGTAAAACTCGTGCCAGC -3:

*reverse*: 5- CATAGTGGGGTATCTAATCCCAGTTTG -3) amplifies an approximately 182 bp fragment conserved across marine fishes and is capable of detecting more than 230 subtropical and temperate species, including the majority of shark and ray taxa. Metabarcoding PCR reactions followed the same master mix composition as above, with an annealing temperature of 60 °C and 35 cycles. These reactions were performed exclusively in the laboratory to ensure contamination control and optimal amplification of low-concentration templates.

All PCR runs included positive controls (white shark and shortfin mako muscle tissue, and aquarium-derived eDNA), negative template controls, and extraction blanks to identify contamination or amplification errors. Amplification success was confirmed via gel electrophoresis prior to sequencing. No contamination was detected in negative controls across field or laboratory workflows.

### 5.2.4 Particle Tracking Simulation

To identify the approximate location of white shark individuals, detected eDNA coordinates were used as seed points for a Lagrangian hindcast simulation using OpenDrift [38] (Figure 5.1F). We modeled particle trajectories with 3-hourly current data from the Copernicus Marine Service (MEDSEA Analysis Forecast [32]) and added wind and current uncertainty terms (0.1 and 0.2—magnitudes around the mean). One thousand particles were released per detection and tracked backwards through 128 hours in 15-minute steps, encompassing the maximum expected lifespan of detectable eDNA in seawater [34]. Particle endpoints were used to infer probable eDNA origin areas.

### 5.2.5 White Shark Assay

Extracted DNA from onboard and laboratory workflows was first quantified using a Qubit Fluorometer v3 (Thermo Fisher Scientific) with a high-sensitivity double-stranded DNA assay to determine total DNA concentration per sample. This quantification step enabled quality control prior to amplification and sequencing and provided an estimate of overall eDNA yield from each filter.

White shark detections were confirmed through Sanger sequencing of amplified CYTB fragments, followed by filtering amplicon variants using the `dada2` pipeline [25]. Representative amplicons were then compared to reference mitochondrial genomes using BLASTn searches against the NCBI nucleotide database [115] to verify taxonomic identity. Sequence alignments were considered positive when match identity exceeded 95% similarity to the white shark reference sequence.

To evaluate potential FP detections resulting from cross-amplification with shortfin mako, we performed a multiple-sequence alignment including reference mitochondrial genomes of both species, a white shark tissue sample, and four positive eDNA samples (Figure 5.3). Single Nucleotide Polymorphisms (SNPs) were recorded to differentiate true white shark signals from mako sequences. Seventeen diagnostic SNPs were identified across the target CYTB fragment, providing consistent separation between species and validating the specificity of the assay [80].

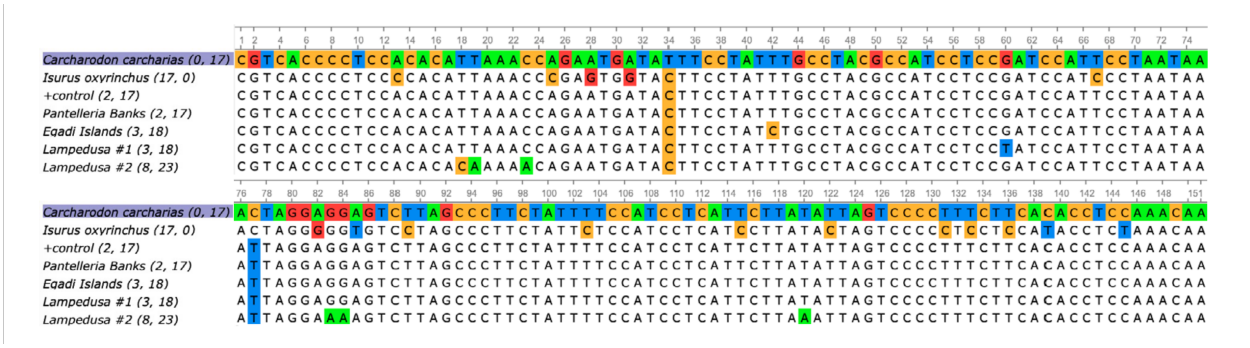


Figure 5.3: Sequence alignment of the white shark (*Carcharodon carcharias*) 151 bp fragment of the mitochondrial CYTB gene. Colored bases indicate either consensus alignment or mismatches relative to the shortfin mako (*Isurus oxyrinchus*) and eDNA sample sequences. Sample labels denote the number of SNPs relative to each species (# SNPs to white shark, # SNPs to shortfin mako). Alignment visualizations were generated using the software Unipro UGENE [156].

### 5.2.6 Metabarcoding

To assess elasmobranch diversity, a subset of 59 environmental samples was selected for metabarcoding analysis. This subset included nine citizen science samples collected from the Adriatic Sea, Sicilian Channel, Tyrrhenian Sea, and Menorca, as well as samples from the 2023 Tunisia to Malta transect containing a confirmed white shark detection used as a positive eDNA control. In addition to environmental samples, the library included three field blanks, one PCR blank, and tissue-derived positive controls from white shark, shortfin mako, porbeagle, and blue shark (*Prionace glauca*). Three aquarium samples with known elasmobranch compositions were also processed to evaluate the taxonomic sensitivity and specificity of the MiFish-E assay by verifying if expected species were detected and whether any non-target amplifications occurred. Observed shark and ray species in aquarium touch tanks included white spotted bamboo shark (*Chiloscyllium plagiosum*), black spotted bamboo shark (*Chiloscyllium punctatum*), chain dogfish (*Scyliorhinus retifer*), dusky smoothhound (*Mustelus canis*), epaulette shark (*Hemiscyllium ocellatum*), yellow stingray (*Urobatis jamaicensis*), bluespotted ribbontail ray (*Taeniura lymma*), zebra shark (*Stegostoma*

*tigrinum*), and Atlantic stingray (*Hypanus sabinus*).

Metabarcoding employed the MiFish-E universal primer set targeting a 182 bp fragment of the mitochondrial 12S rRNA gene [109]. Amplification was conducted following a two-step PCR protocol to incorporate dual indexing for sample multiplexing. Libraries were prepared using Nextera XT v2 Set A indices and sequenced on an Illumina MiSeq Micro V2 300-cycle flow cell ($2 \times 150$ bp) at the VT-GSC, generating approximately 4 million paired-end reads per lane [46]. Library preparation followed the Illumina *GenerateFASTQ* workflow. Each library batch incorporated 10% PhiX control to balance base diversity.

### 5.2.7 Taxonomic Assignment

Bioinformatic processing of metabarcoded sequences followed a standardized workflow implemented in R using the `dada2` [25] and `DECIPHER` [173] packages. Raw paired-end reads were quality-filtered and trimmed to infer amplicon sequence variants (ASVs) with `dada2`, removing low-quality sequences. High-confidence ASVs were compared against a curated elasmobranch mitochondrial reference database derived from 12S rRNA gene sequences using `blastn` [115]. Taxonomic assignments were accepted when sequence identity exceeded 90% and alignment coverage surpassed 80% of the target seauence [161]. Relative abundance was calculated per sample, and ASVs representing less than 1% of total reads were excluded to minimize noise from sequencing cross-contamination [161]. This workflow enabled robust identification of elasmobranch taxa from eDNA while controlling for FP detections through the use of filtering thresholds and the use of reference genes specific to sharks and rays.

## 5.3 Results

### 5.3.1 White Shark Assay

**Collecting Samples.** Between 2021 and 2024, a total of 204 eDNA samples were collected across the Sicilian Channel, Gulf of Gabès, Malta Plateau, Tyrrhenian and Ligurian Seas, Adriatic Sea, Ionian Sea, and Menorca (Figure 5.7). Sampling depths ranged from the surface to 100 m, with an average depth of $21.6 \pm 20.5$ m and an average filtered volume of $2 \pm 1$ liters.

**Detecting White Sharks.** White shark DNA was detected in four samples from three stations during the 2021 expedition, corresponding to sites near the Egadi Islands, Pantelleria Banks, and Lampedusa. An additional detection occurred near Lampedusa in 2023, first confirmed with TapeStation (Figure 5.4). These detections were verified through Sanger sequencing and BLASTn comparison [115] against reference mitochondrial genomes in the NCBI database, each showing greater than 95% identity to the white shark CYTB sequence.

**Particle Simulation.** Hindcast particle simulations from 2023 fieldwork (see Figure 5.5, 54, 80) identified probable eDNA source regions within 48–128 h prior to detection. This modeling effort led to the visual observation of an adult female shortfin mako near Lampedusa in 2023 (see Figure 5.6), emphasizing both the potential of eDNA-informed field strategies and the necessity of improving assay specificity to white sharks.

**Assay Sensitivity.** To address possible cross-amplification between the two lamnid species, multiple-sequence alignment of reference genomes, tissue-derived samples, and four positive eDNA detections differed by an average of only $4 \pm 2.7$ nucleotide positions from the white

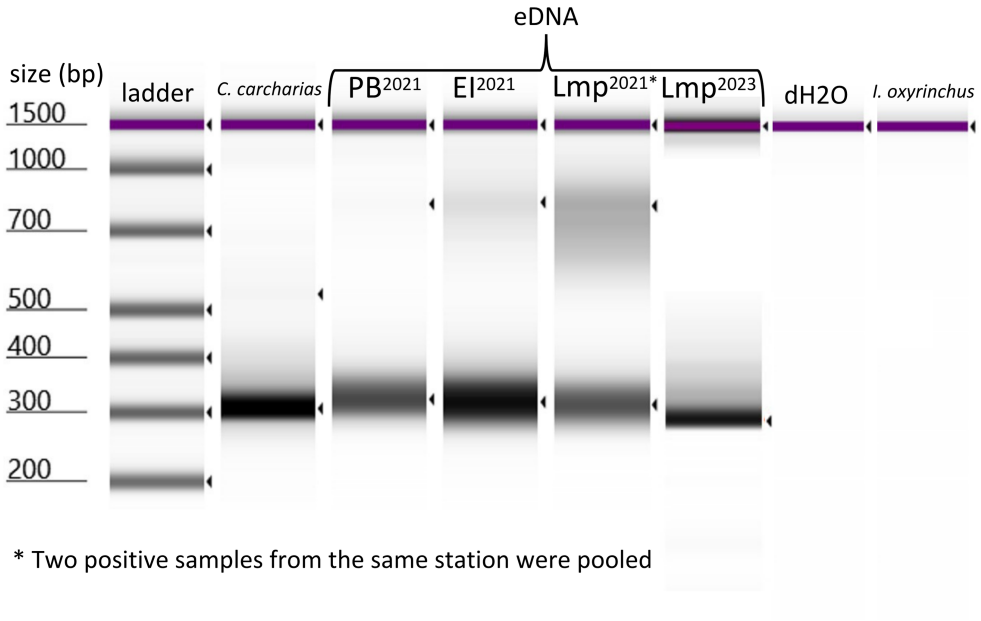


Figure 5.4: Electrophoresis results confirming eDNA amplification of the white shark (*Carcharodon carcharias*). Samples are labeled by station: PB – Pantelleria Banks, EI – Egadi Islands, and Lmp – Lampedusa. California white shark tissue served as the positive control, distilled water as the negative control during PCR, and Mediterranean shortfin mako (*Isurus oxyrinchus*) tissue as an FP indicator. Two white shark detections from the same Lampedusa station in 2021 were pooled for electrophoresis and sequencing.

shark reference genome, compared to 19±2.7 differences from the shortfin mako (Figure 5.3). This indicated that the eDNA sequences were approximately 79% more likely to represent white shark detections than mako.

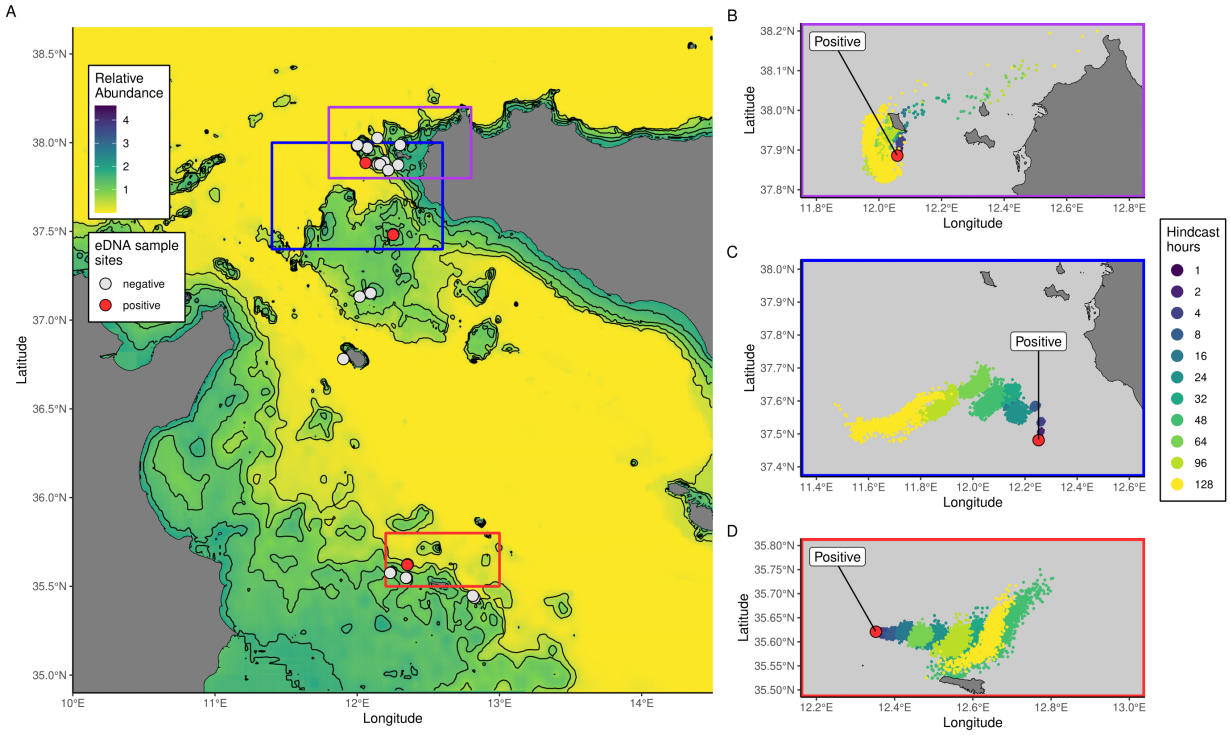


Figure 5.5: Predicted relative abundance and particle dispersal of white shark (*Carcharodon carcharias*) eDNA in the Sicilian Channel. (a) Modelled relative abundance of white sharks during May–June. (b–d) Lagrangian particle tracking hindcasts showing predicted locations of white shark eDNA molecules prior to detection. In hindcasted hours, purple represents the most recent predicted locations (1 hour prior to detection), while yellow indicates positions 128 h prior. Red markers denote sampling stations where white shark eDNA was detected.

### 5.3.2 Elasmobranch Detections

**Biodiversity.** Preliminary metabarcoding of 48 eDNA samples detected 12 elasmobranch species with >95% match identity and >1% relative abundance of ASVs (Table 5.1). At least one species was detected in 37 samples. Species richness per sample ranged from 1

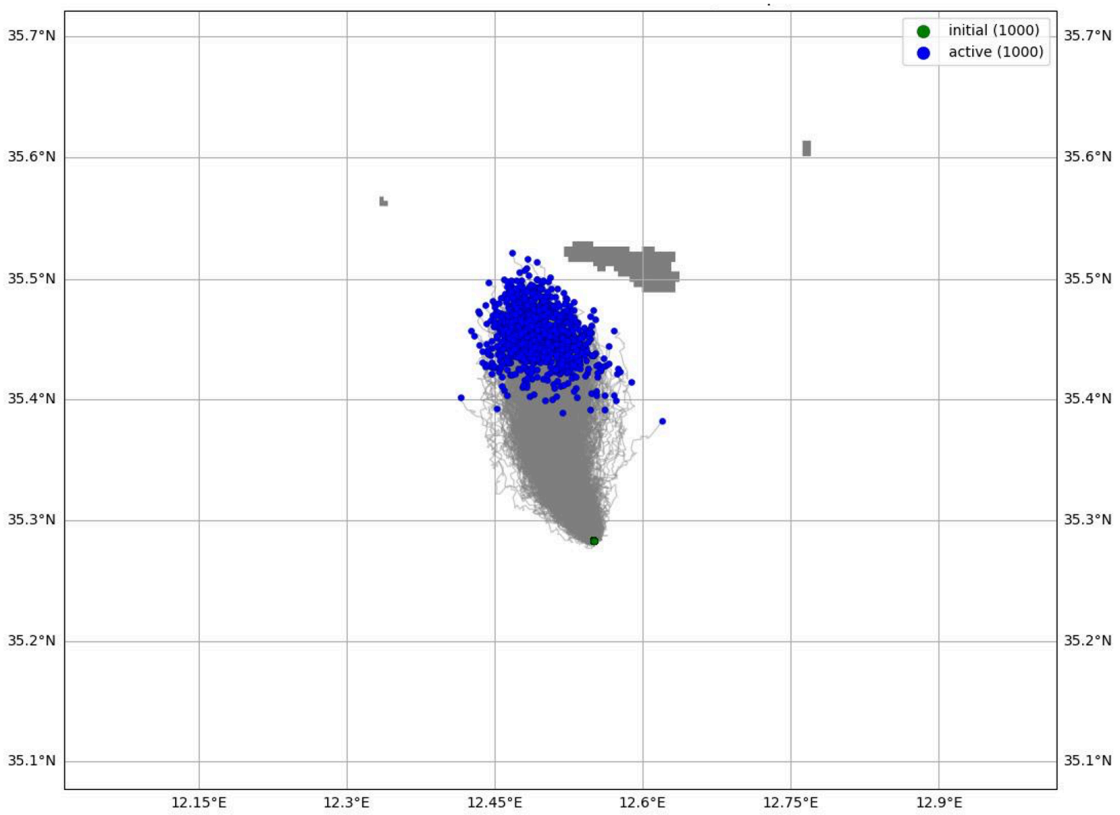


Figure 5.6: Hindcasted eDNA particle dispersal 48 hours from the time of detection and coordinates of the white shark detection in 2023, south of Lampedusa. The green point represents the seeded particles at the detection coordinate, and the blue points represent the backward-dispersed particle simulation.

to 9 taxa, with a mean richness of 2.2 shark species per sample. These results, though preliminary, demonstrate the sensitivity of metabarcoding for capturing multi-species signals from seawater samples.

| Species | Common name | Basins detected | Relative abundance (%) |
|---|---|---|---|
| *Isurus oxyrinchus* | Shortfin mako | Adriatic Sea, Ligurian Sea, Sicilian Channel | 54.05 |
| *Carcharodon carcharias* | White shark | Adriatic Sea, Sicilian Channel | 35.14 |
| *Lamna nasus* | Porbeagle shark | Adriatic Sea, Tunisian Plateau | 13.51 |
| *Carcharhinus brachyurus* | Copper shark | Adriatic Sea | 10.81 |
| *Prionace glauca* | Blue shark | Adriatic Sea | 10.81 |
| *Carcharhinus falciformis* | Silky shark | Adriatic Sea | 8.11 |
| *Carcharhinus obscurus* | Dusky shark | Adriatic Sea | 8.11 |
| *Carcharhinus plumbeus* | Sandbar shark | Adriatic Sea | 8.11 |
| *Carcharhinus albimarginatus* | Silvertip shark | Adriatic Sea | 8.11 |
| *Taeniura lymma* | Bluespotted ribbontail ray | Adriatic Sea | 8.11 |
| *Rhinobatos cemiculus* | Blackchin guitarfish | Ligurian Sea | 2.70 |
| *Chiloscyllium plagiosum* | Whitespotted bamboo shark | Adriatic Sea | 2.70 |

Table 5.1: Summary of 12 elasmobranch species detected across Mediterranean basins from 37 eDNA samples (2021–2024). The table lists each species, corresponding common name, the number of detections, the basins where they were found, and their relative abundance among total samples.

**Important Detections.** Across all samples, the most frequently detected species were the shortfin mako (*Isurus oxyrinchus*: 20 samples), white shark (*Carcharodon carcharias*: 13 samples), and porbeagle (*Lamna nasus*: 5 samples) (Table 5.1 and Figure 5.7). Blue shark (*Prionace glauca*) and copper shark (*Carcharhinus brachyurus*) were detected in 4 samples. All three Lamnid species were detected in the Ligurian Sea, Sicilian Channel (including the Tunisian Plateau), and Adriatic Sea, indicating broad regional overlap of critically endangered pelagic shark assemblages despite historical population collapses.

Continued validation and expansion of the reference database will improve discrimination among closely related taxa and reduce uncertainty. Nevertheless, these early findings

underscore the promise of molecular community assays for reconstructing elasmobranch biodiversity in the Mediterranean Sea and identifying remaining hotspots of Lamnid occurrence.

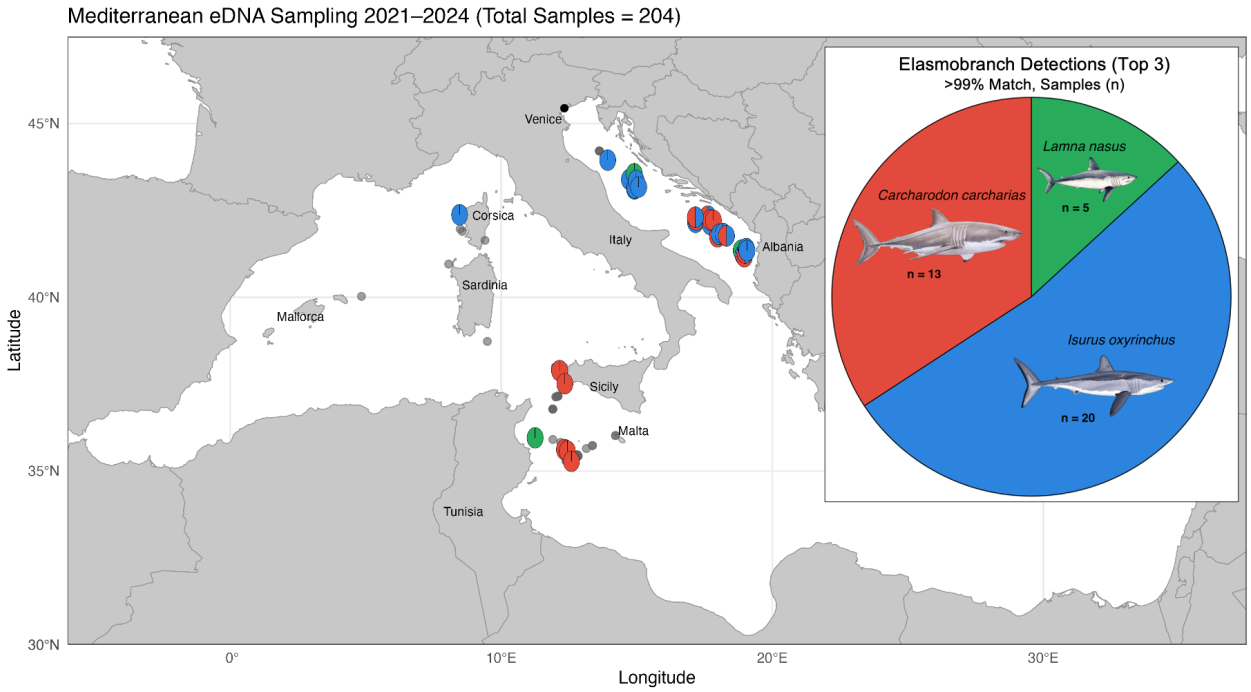


Figure 5.7: Spatial distribution of all eDNA sampling stations across the Mediterranean Sea from 2021–2024. Colored markers denote sample-specific species detections, with the inset pie chart summarizing the detection of the three Lamnid shark species: white shark (*Carcharodon carcharias*, red), shortfin mako (*Isurus oxyrinchus*, blue), and porbeagle (*Lamna nasus*, green).

## 5.4 Discussion

Our integrated eDNA frameworks revealed the presence of critically endangered apex predators in one of the world's most data-poor seas. Species-specific and multi-species assays combined with field and laboratory workflows, and particle tracking models allowed us to detect and track white sharks, while characterizing biodiversity and relative detection abundance across multiple years and Mediterranean regions [94, 96, 161]. These results

confirm earlier reports of remnant Mediterranean populations of white sharks [63, 96, 111] and show that molecular tools can supplement conventional observation in regions where encountering these animals is rare [111, 118]. We show elasmobranch biodiversity, particularly large sharks, persisting across the Sicilian Channel and Tunisian Plateau, Tyrrhenian and Ligurian Sea, and Adriatic Sea despite long-term declines. Together, these outcomes strengthen occurrence baselines for threatened elasmobranchs and inform spatial prioritization for conservation.

A major contribution of this work is demonstrating how targeted eDNA assays serve roles complementary to conservation needs. White shark markers provided confident genetic detections and mitigated the risk of misclassification with the genetically similar species shortfin mako. These results provided transparent spatiotemporal distributions of presence for the critically endangered species, guiding future efforts to observe them [54]. In parallel, biodiversity assays captured broader elasmobranch assemblages, highlighting hotspots congruent with historical observations [18, 42, 44, 50, 58, 134, 135, 147, 165]. However these results should be interpreted cautiously, as further bioinformatic refinement is required to incorporate negative controls, field blanks, and aquarium reference samples for standardization. Species that are not native to the Mediterranean are likely misidentifications (see Table 5.1) such as the bluespotted ribbontail ray (*Taeniura lymma*) and whitespotted bamboo shark (*Chiloscyllium plagiosum*) both primarily found in the Indo-West Pacific [93, 137].

These assays addressed both the immediacy of locating white sharks and the longer-term need to monitor shifting community structure under climate and fishing pressures [42]. The widespread detections and high relative abundance of critically endangered Lamnids, while preliminary, provide rare insight into a population that may be showing early signs of recovery, in contrast to smaller demersal sharks that were strikingly absent from results, possibly due to the lack of reference libraries describing them.

Furthermore, citizen sampling offered scalability, provided participators were adequately instructed to handle contamination-prone samples. Our sampling network detected two critically endangered species: the shortfin mako, and notably the blackchin guitarfish (*Rhinobatos cemiculus*) in a region of historical abundance but recent decline. The Mediterranean International Trawl Surveys (MEDITS) reported that the species has largely disappeared from the northern Mediterranean [92], underscoring the potential of citizen-collected eDNA samples to reveal species otherwise missed by conventional surveys.

The integration of molecular detection with particle tracking models creates a dynamic decision-support tool for research expeditions. Rapid onboard screening enabled near-real-time awareness (about 4.5 hours) of species presence, while hindcasts identified likely origin areas of detected eDNA. This approach, following methods in Andruszkiewicz et al. [4] and Dagestad et al. [38], contextualizes detections within physical oceanography, improving inference about animal movement and enhancing efficiency of follow-up surveys. The 2023 detection and subsequent sighting of a shortfin mako near Lampedusa underscore both the promise and the challenge of assay specificity, highlighting the need for continued refinement of molecular markers for rare species in mixed pelagic systems. Strengthening assay performance will require targeted collection of region-specific tissue samples from priority taxa such as the white shark and shortfin mako to build a more representative regional reference library [58, 96]. Sequencing and bioinformatically comparing these genomes across mitochondrial and nuclear loci will enable the design of primers optimized for local haplotypes, minimizing the risk of cross-amplification and false detections [94, 161]. Refining a species-specific workflow to additionally become region-specific is the next step to improve diagnostic accuracy and ensure that detections reflect true presence rather than genetic similarity within sympatric Lamnid populations.

Despite these advancements, limitations remain inherent to marine eDNA. Its high sensi-

tivity increases the risk of contamination-related FP detections, while environmental degradation may result in missing detections [34]. We mitigated these risks through field blanks, tissue and aquarium controls, and conservative sequence identity thresholds, yet the full incorporation of negative controls and more region-specific genetic materials remains a future priority. Primer bias and incomplete reference databases can distort representation of taxa, issues acknowledged in global metabarcoding efforts [109]. White shark detections identified by the species-specific assay were independently confirmed in the multi-species metabarcoding assay, reinforcing the reliability of both methods.

To overcome current limitations in eDNA resolution and the broader genetic characterization of Mediterranean elasmobranchs, we are systematically monitoring fisheries landings across Tunisian ports and have collected an unprecedented repository of verified tissue samples. This regional genetic archive provides the foundation for developing more sensitive species assays and enabling population-level analyses. From these data, forthcoming studies will estimate effective population size, genetic diversity, and long-term connectivity. These are key parameters for understanding demographic resilience and validating the results presented here.

Future efforts should aim to quantitatively validate how well these assays detect target species by predicting detection likelihood and comparing results directly with traditional monitoring methods such as BRUVs and fisheries catch and tagging records (e.g., MED-ITS and Shea et al. [135]). These findings support the emerging hypothesis that eDNA assays can reliably detect rare and elusive shark species in the Mediterranean, and that their detection probabilities are expected to scale predictably with indicators from conventional survey approaches. Testing this relationship will be essential for integrating molecular and observational datasets into unified monitoring frameworks that can better estimate species occurrence and abundance.

Verified white shark detections and initial multi-species co-occurrence results across major regional basins highlight remnant hotspots and migratory corridors that warrant heightened protection. These molecular baselines can refine species distribution models, support IUCN assessments, and inform management across Mediterranean nations. The methodological blueprints established here-integrating field and laboratory protocols, taxa-specific assays, ocean modeling, and citizen engagement—provides a reproducible framework for eDNA-based monitoring in other data-poor marine systems. As reference libraries and analytical precision increases, eDNA approaches will increasingly transition from exploratory research to operational monitoring that guides the conservation of global shark populations.

# Chapter 6

# Conclusions

## 6.1 Main Conclusions

In summary, my thesis set out to (i) develop and validate deep learning models for automatically detecting and classifying visual media of sharks to the species level, (ii) aggregate, clean, and predict shark observations from major social networks and online platforms to estimate relative abundance and assess temporal trends across case-study regions, and (iii) apply species- and community-level Environmental DNA (eDNA) assays in parallel with oceanographic modeling and citizen science sampling, for mapping the occurrence of threatened Mediterranean shark populations. In this chapter, I summarize the main conclusions of these three objectives, and highlight the practical conservation and management implications of my findings, then recommend immediate and long-term pathways for building on the scientific potential of this dissertation.

The findings presented in Chapter 2 demonstrate the best performing (to date) automatic detection and classification tool for filtering for visual media of sharks. The Shark Detector (SD) effectively filtered and classified heterogeneous media from Instagram (IG), Baited Remote Underwater Video Systems (BRUVs), and online videos, and outperformed iNaturalist (iNat)'s wildlife classification Artificial Intelligence (AI) *Seek* on a sample of 400 random shark images (73% vs. 62% respectively). SD versions 1–5 build upon the frameworks presented in Chapters 2–4, incorporating new training data and iterative improvements in

performance, speed, and model architecture. The SD achieved strong end-to-end performance (detection 89%, binary shark vs. non-shark filtering 91%, species classification 80%). Predicting accuracy vs. training data quantity informed crucial data thresholds at the genus and species taxonomic ranks. These findings highlighted impacts to the performance such as the diversity of taxonomic classes and confusion at the order → family → genus → species levels, morphological distinction, and image data quality and quantity. These findings continue to guide how to optimize training data balance for iterative SD versions. Data-poor and morphologically similar taxa confused the model, underscoring the value of continued taxa balancing and morphological diversity.

In Chapter 3, I increased the speed and performance of the SD and expanded its accessibility for practical field-based operations to post-process 14 hours of BRUV footage from biodiversity surveys in Hawaii and Palau. With SharkByte, I developed a Graphical User Interface (GUI) tool and detected and classified >45 sharks and 7 species with 94% accuracy, speeding up video annotation by up to 95%– generating species richness and abundance indices. The R package `sharkDetectoR` and companion Application Programming Interface (API) successfully classified over 46k sharkPulse images of 80 species in 16 minutes with 92% species-specific accuracy. These findings bridged the gap between large-scale image and video data processing and varying levels of user computational expertise.

Chapter 4 applied stepwise workflows of data science, AI, and statistical inference to source, filter, and sanitize observations of sharks from four social network (SN) platforms and online archives, and subsequently predict temporal trends of relative abundance. I crowdsourced 5.4 million raw posts and 700k unique shark observations from IG, Flickr, iNat, and the Global Biodiversity Information Facility (GBIF). From Flickr and iNat, I additionally crowdsourced proxies of user observation effort to standardize sightings. By assuming a negative binomial distribution of observations within a generalized linear modeling framework, I

fit relative abundance trends for 17 shark species, both large pelagic species and coastal reef-associated species, in the Bahamas and Hawaii regions. I chose these regions because they had historical assessments to validate SN trends with independent surveys of relative abundance including BRUVs, and scientific fishing and scuba diving operations. I found that iNat trends significantly aligned with previous assessments in both regions, while Flickr trends reflected stronger uncertainties. Incorporating AI-generated observations without human verification contributed to these uncertainties, as well as stronger evidence of observation bias towards large, charismatic species. Overall, I showed that SN trends provide a practical, low-cost complement to traditional monitoring with global application. This approach enables early detection of declines or recoveries and creates scalable, data-driven frameworks for integrating citizen observations into formal management and policy decisions.

Chapter 5 detects the eDNA of white sharks that have become exceptionally rare in the Mediterranean Sea due to overfishing, habitat degradation, and increasingly warming seas. In 4.5 hours from collecting seawater, I detected white shark eDNA at four sampling stations throughout the Sicilian channel, in 2021 and 2023, while onboard expedition vessels. The detections were later confirmed in the laboratory with Sanger sequencing, matching sequences with reference databases through `blatn` commands, and multiple-sequence alignment. Tracking particles with oceanographic simulations revealed eDNA shedding within 12 nautical miles of where the animal was detected, indicating directional and temporal presence of the animal. In 2023, particle tracking results guided the targeted deployment of BRUVs near the eDNA detection site, which directly led to the visual confirmation of an adult female shortfin mako. The white shark assay showed a 21% False Positive (FP) rate—about one in five detections could be misidentified—but all detections so far were confirmed true, underscoring the importance of the laboratory sensitivity workflows demonstrated here. From 2023–2024, I expanded this study with a new sampling transect across

the Adriatic Sea and launched a citizen science initiative for engaging Mediterranean sailors with easy-to-use sampling kits. This effort broadened geographic coverage of eDNA monitoring. Subsequently, through an opportunistic elasmobranch-specific metabarcoding study, I characterized biodiversity of new and previous samples, detecting 12 species in the Adriadic and Ionian Seas, and the Northwestern basins of the Mediterranean Sea. From the findings, the three most frequently detected species are the region's most threatened Lamnids. All samples yielding white shark detections with the targeted assay were likewise positive with the metabarcoding assay, demonstrating the latter's sensitivity.

## 6.2 Implications to Management and Education

The collective findings of this dissertation highlight the potential of emerging digital and molecular technologies for advancing shark conservation and management. Automated image classification, big data analytics, and molecular detection address a persistent barrier in conventional monitoring: the lack of standardized, recent, and species-specific information at meaningful ecological scales. Together, these tools demonstrate how to bridge that information gap and deliver reproducible data streams. This section highlights the key management implications of my findings, as well as education and outreach achievements, and explains how these standardized and validated approaches can be embedded within current frameworks for shark monitoring and population assessment.

The Chapters 2–3, and published work Jenrette et al. [78], Varini et al. [164], and Jenrette et al. [75] shows how deep learning frameworks can localize and construct ecologically relevant occurrence records from massive, heterogeneous, and noisy visual datasets that had never before been applied to shark monitoring at this scale. The resulting capacity and cost-efficiency to automatically detect, classify, and quantify sharks from visual media has

immediate management value especially when resources are limited. Accessibility of these detection and classification functions through the SharkByte GUI and `sharkDetectoR` package empowers a broader range of users—including field researchers, non-governmental organizations, and citizen scientists—to generate standardized occurrence records from systematic (e.g., BRUVs) and opportunistic (e.g., social networks, online archives, user submissions) surveys.

The methods outlined in Chapter 4, Ferretti et al. [52], and Jenrette et al. [76] collect and model opportunistic encounters generated by a broad demographic of ocean users, producing spatially and temporally rich data. This diversity of users enables the detection of ecological change in regions and time periods where professional monitoring is logistically or financially challenged. These records and modeled trends can be directly incorporated into national and regional assessment organizations such as the International Union for Conservation of Nature (IUCN) and National Oceanic and Atmospheric Administration (NOAA), complementing fishery-dependent reporting systems and strengthening the empirical basis for Red List assessments and stock evaluations. The generation of standardized Sightings per Unit Effort (SPUE) indices for 17 species in both managed regions (the Bahamas) and more vulnerable areas (Hawaii) provide empirical baselines for evaluating the effectiveness of existing protective measures such as Marine Protected Area (MPA)s, shark sanctuaries, and gear restrictions. They also enable managers to interpret population trends within the context of human presence and pressure—where increased digital sightings may reflect both rising observation effort from urbanized coasts and heightened exploitation risks of increased fishing and ocean use pressure—highlighting the need for management actions that distinguish ecological recovery from intensified human activity.

In Chapter 5, Jenrette et al. [80], and Ferretti et al. [54], molecular forensics provided crucial detections of critically endangered Mediterranean shark populations, validating the

power of eDNA to reveal species presence in regions where visual encounters are extremely rare. The integration of oceanographic particle tracking, laboratory verification, and targeted BRUV deployments establishes a practical workflow for adaptive sampling and validation. This framework could be adopted by regional fisheries bodies and environmental agencies to prioritize sampling zones, evaluate the effectiveness of MPAs, and verify the persistence of threatened populations through continuous, non-invasive surveys. The citizen science extension of the eDNA program further demonstrates how participatory monitoring can expand geographic coverage, engage stakeholders, and promote public stewardship of marine resources.

Effective conservation requires not only scientific innovation but also active communication, collaboration, and public engagement to translate research into management action. To this end, I have prioritized outreach and education throughout this dissertation to maximize the accessibility and real-world impact of the technologies developed herein. During ongoing field operations in Hawaii, I presented the value and implementation of the SD in a webinar hosted by the Ocean Exploration Trust and National Geographic [79], demonstrating to managers, researchers, and the public how AI can be directly integrated into shark biodiversity assessments. Through a series of sharkPulse hackathon events [77, 90], I advanced the platform toward a synergistic crowdsourcing and citizen science verification system—an essential step for enhancing data resolution and public engagement in conservation outcomes. In partnership with Oxford and Imperial College London, I expanded automatic shark object-detection programs [164] that have many applications including electronic monitoring systems used by commercial fisheries, offering a scalable means of bycatch verification.

To broaden the dissemination of molecular approaches, I conducted interviews with *Forbes* [102] and the SeaKeeper's Society and Virginia Tech Genomics Sequencing Center (VT-GSC)

[143] to communicate the conservation status of Mediterranean sharks and the unique role of eDNA in detecting populations that elude conventional monitoring. In collaboration with the SeaKeeper's Society, I helped establish a citizen network of sailing vessels that now contributes eDNA samples across the Mediterranean Sea—an initiative that simultaneously expands scientific coverage and strengthens stakeholder participation in regional conservation. Collectively, I trust these efforts as models for applied conservation science, where the dissemination of tools, data, and knowledge empowers both management institutions and the public to contribute directly to transparent, data-driven decision-making.

## 6.3 Future Research

The future of non-invasive, cost-efficient wildlife monitoring lies in the continued development and integration of digital and molecular data-handling mechanisms (Appendix A). The research presented in this dissertation establishes a foundation for such systems, but several key directions should now be prioritized to extend their scientific, operational, and management impact. These efforts will further bridge the gap between innovation and application, transforming automated and molecular observations into actionable conservation tools.

1. **Building human–machine validation networks:** The SD and its associated platforms demonstrate that data quantity is no longer the limiting factor in digital monitoring, data quality is. While millions of shark images are now available, the next challenge is developing streamlined verification systems that balance automation with human oversight. Future research should focus on integrating sharkPulse with a network of trained citizen scientists who can validate taxonomy, location, and temporal metadata through incentive-based interfaces such as in Horn et al. [69], Sullivan et al.

[149]. We have initiated this system on the sharkPulse platform in Jenrette et al. [77] and Kothari et al. [90], but more backend programmatic work needs to be done to gain a consistent citizen audience. In parallel, the SD continues to evolve as both a filtering and species-classification framework, and other machine-learning models should be trained to flag duplicates, reposts, and range anomalies, creating a feedback loop in which validated human inputs continuously improve automatic outputs. Establishing this iterative verification architecture will replace tedious manual review with near-perfect automatic annotations [69] and produce a globally scalable, quality-assured database for management use.

2. **Adapting automated detection for fisheries monitoring:** Novel shark detection software can immediately address one of the most persistent obstacles in shark conservation, the lack of high quality catch data described to the species level [15, 44]. Future work should integrate automated object detection into electronic monitoring systems aboard commercial and artisanal fishing vessels. A trained model capable of reliably identifying and classifying shark bycatch events would provide real-time species indices [162] that can be deployed globally to reduce observer burden, improve logbook accuracy, and generate standardized catch data for stock assessments [23, 130]. The SharkByte platform offers a promising foundation for such integration: its next phase should involve field testing under live fishing conditions to evaluate precision, processing speed, and usability within existing electronic monitoring frameworks operated by NOAA and partner management agencies.

3. **Scaling digital observation data for global population assessments:** Social networks and online archives will continue to grow as dominant sources of wildlife observations, providing unprecedented coverage across time and geography [159]. Future research should expand programmatic efforts to stabilize, standardize, and central-

ize these datasets into near–real-time monitoring systems. The goal is to democratize ecological data collection by transforming opportunistic human encounters into structured, quantitative indicators of population health. Data science and statistical approaches should be refined to build platform-specific models that account for bias, validate population trends against conventional surveys, and transfer standardized outputs directly to management and conservation bodies such as NOAA and the IUCN, or international conservation frameworks such as the Convention on Migratory Species. Semi-automatically identifying and tracking digital footprints for revealing relative abundance and distribution patterns reflects human presence, exploitation intensity, and conservation success with a new perspective.

4. **Expanding molecular forensics and international collaboration:** Accurately identifying molecular footprints of marine fauna represents a revolutionary advance in non-invasive monitoring, and continued research in this area is critical. Future directions should include expanding eDNA sampling across the Mediterranean and adjacent basins, developing more sensitive species- and population-level assays, and strengthening bioinformatic pipelines for detection validation. Building a comprehensive genetic reference database for Mediterranean elasmobranchs will substantially improve molecular diagnostics [110]. This should be pursued through collaborative sequencing of tissue samples collected from port-monitoring activities across the Mediterranean. Current efforts in Tunisia set an example of establishing consistent monitoring and biological sampling operations in low-capacity regions common throughout the Mediterranean regions. Importantly, such research must be implemented through equitable international partnerships that empower local scientists and institutions to lead conservation genomics within their jurisdictions. Collaborative molecular research in countries facing strong fishing pressures will not only improve biodiversity monitoring but also build

capacity in regions most critical to shark conservation.

# Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR*, abs/1603.04467, 2016. doi: 10.48550/arXiv.1603.04467.

[2] Vaneeda Allken, Nils Olav Handegard, Shale Rosen, Tiffanie Schreyeck, Thomas Mahiout, and Ketil Malde. Fish species identification using a convolutional neural network trained on synthetic data. *ICES Journal of Marine Science*, 76:342–349, 10 2018. ISSN 1054-3139. doi: 10.1093/icesjms/fsy147.

[3] Elizabeth A Andruszkiewicz, Lauren M Sassoubre, and Alexandria B Boehm. Persistence of marine fish environmental DNA and the influence of sunlight. *PLOS ONE*, 12: 1–18, 9 2017. doi: 10.1371/journal.pone.0185043. Publisher: Public Library of Science.

[4] Elizabeth A Andruszkiewicz, Jeffrey R Koseff, Oliver B Fringer, Nicholas T Ouellette, Anna B Lowe, Christopher A Edwards, and Alexandria B Boehm. Modeling environmental DNA transport in the coastal ocean using Lagrangian particle tracking. *Frontiers in Marine Science*, page 477, 2019. doi: 10.3389/fmars.2019.00477.

[5] Jacob Asher. *A Deeper Look at Hawaiian Coral Reef Fish Assemblages: A Comparison of Survey Approaches and Assessments of Shallow to Mesophotic Communities*. PhD thesis, 2017. URL http://hdl.handle.net/20.500.11937/59686. PhD Thesis.

[6] Jacob Asher, Ivor D Williams, and Euan S Harvey. An Assessment of Mobile Predator Populations along Shallow and Mesophotic Depth Gradients in the Hawaiian Archipelago. *Scientific Reports*, 7:3905, 6 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-03568-1. URL https://doi.org/10.1038/s41598-017-03568-1.

[7] Jacob Asher, Ivor D Williams, and Euan S Harvey. Mesophotic Depth Gradients Impact Reef Fish Assemblage Composition and Functional Group Partitioning in the Main Hawaiian Islands. *Frontiers in Marine Science*, Volume 4 - 2017, 2017. ISSN 2296-7745. URL https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2017.00098.

[8] Jacob Asher, Ivor D Williams, and Euan S Harvey. Is seeing believing? Diver and video-based censuses reveal inconsistencies in roving predator estimates between regions. *Marine Ecology Progress Series*, 630:115–136, 2019. ISSN 01718630, 16161599. URL https://www.jstor.org/stable/26920544.

[9] Monica Barone, Frederik H Mollen, Jenny L Giles, Lindsay J Marshall, Melany Villate-Moreno, Carlotta Mazzoldi, Elisa Pérez-Costas, Jürgen Heine, and Cástor Guisande. Performance of iSharkFin in the identification of wet dorsal fins from priority shark species. *Ecological Informatics*, 68:101514, 2022. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2021.101514.

[10] David C Bartholomew, Jeffrey C Mangel, Joanna Alfaro-Shigueto, Sergio Pingo, Astrid Jimenez, and Brendan J Godley. Remote electronic monitoring as a potential alternative to on-board observers in small-scale fisheries. *Biological Conserva-*

*tion*, 219:35–45, 2018. ISSN 0006-3207. doi: 10.1016/j.biocon.2018.01.003. URL https://www.sciencedirect.com/science/article/pii/S0006320717307899.

[11] Vijay Barve and Edmund Hart. ***rinat**: Access 'iNaturalist' Data Through APIs*, 2022. URL https://CRAN.R-project.org/package=rinat. R package version 0.1.9.

[12] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30:89–116, 2015. ISSN 1435-5655. doi: 10.1007/s00146-014-0549-4. URL https://doi.org/10.1007/s00146-014-0549-4.

[13] Julia K Baum and Wade Blanchard. Inferring shark population trends from generalized linear mixed models of pelagic longline catch and effort data. *Fisheries Research*, 102: 229–239, 2010. ISSN 0165-7836. doi: 10.1016/j.fishres.2009.11.006.

[14] Julia K Baum and Ransom A Myers. Shifting baselines and the decline of pelagic sharks in the Gulf of Mexico. *Ecology Letters*, 7:135–145, 2004. doi: 10.1111/j.1461-0248.2003.00564.x. URL https://doi.org/10.1111/j.1461-0248.2003.00564.x.

[15] Julia K Baum, Ransom A Myers, Daniel G Kehler, Boris Worm, Shelton J Harley, and Penny A Doherty. Collapse and Conservation of Shark Populations in the Northwest Atlantic. *Science*, 299:389–392, 2003. ISSN 0036-8075. URL https://science.sciencemag.org/content/299/5605/389.

[16] Kingsly C. Beng and Richard T. Corlett. Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. *Biodiversity and Conservation*, 29(7):2089–2121, 2020. ISSN 1572-9710. doi: 10.1007/s10531-020-01980-0. URL https://doi.org/10.1007/s10531-020-01980-0.

[17] C Boettiger, D T Lang, and P C Wainwright. `rfishbase`: exploring, manipulating

and visualizing FishBase data from R. *Journal of Fish Biology*, 81:2030–2039, 2012. ISSN 1095-8649. doi: 10.1111/j.1095-8649.2012.03464.x.

[18] G Boldrocchi, J Kiszka, S Purkis, T Storai, L Zinzula, and D Burkholder. Distribution, ecology, and status of the white shark, Carcharodon carcharias, in the Mediterranean Sea. *Reviews in Fish Biology and Fisheries*, 27:515–534, 9 2017. ISSN 1573-5184. doi: 10.1007/s11160-017-9470-5. URL https://doi.org/10.1007/s11160-017-9470-5.

[19] Emilie Boulanger, Nicolas Loiseau, Alice Valentini, Véronique Arnal, Pierre Boissery, Tony Dejean, Julie Deter, Nacim Guellati, Florian Holon, Jean-Baptiste Juhel, Philippe Lenfant, Stéphanie Manel, and David Mouillot. Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves. *Proceedings of the Royal Society B: Biological Sciences*, 288: 20210112, 2021. doi: 10.1098/rspb.2021.0112.

[20] Olav Brautaset, Anders Ueland Waldeland, Espen Johnsen, Ketil Malde, Line Eikvil, Arnt-Børre Salberg, and Nils Olav Handegard. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. *ICES Journal of Marine Science*, 77:1391–1400, 1 2020. ISSN 1054-3139. doi: 10.1093/icesjms/fsz235.

[21] John S Bridle. Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In Françoise Fogelman Soulié and Jeanny Hérault, editors, *Neurocomputing*, pages 227–236. Springer Berlin Heidelberg, 1990. ISBN 978-3-642-76153-9. doi: 10.1007/978-3-642-76153-9.

[22] E Brooks, K Sloman, D Sims, and A Danylchuk. Validating the use of baited remote underwater video surveys for assessing the diversity, distribution and abundance of sharks in the Bahamas. *Endangered Species Research*, 13:231–243, 2011.

[23] Christopher J. Brown, Amelia Desbiens, Max D. Campbell, Edward T. Game, Eric Gilman, Richard J. Hamilton, Craig Heberer, David Itano, and Kydd Pollock. Electronic monitoring for improved accountability in Western Pacific tuna longline fisheries. *Marine Policy*, 132:104664, 2021. ISSN 0308-597X. doi: 10.1016/j.marpol.2021.104664. URL https://www.sciencedirect.com/science/article/pii/S0308597X2100275X.

[24] Stephan Bruns and Aaron C Henderson. A baited remote underwater video system (BRUVS) assessment of elasmobranch diversity and abundance on the eastern Caicos Bank (Turks and Caicos Islands); an environment in transition. *Environmental Biology of Fishes*, 103:1001–1012, 2020. ISSN 1573-5133. doi: 10.1007/s10641-020-01004-4. URL https://doi.org/10.1007/s10641-020-01004-4.

[25] B J Callahan, P J McMurdie, M J Rosen, A W Han, A J A Johnson, and S P Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*, 13:581–583, 2016. doi: 10.1038/nmeth.3869. Publisher: Nature Methods.

[26] Steven Campana, Warren Joyce, and Michael Manning. Bycatch and discard mortality in commercially caught blue sharks Prionace glauca assessed using archival satellite pop-up tags. *Marine Ecology-progress Series - MAR ECOL-PROGR SER*, 387:241–253, 7 2009. doi: 10.3354/meps08109.

[27] M Cappo, Euan Harvey, and Mark Shortis. Counting and measuring fish with baited video techniques-an overview. *AFSB Conference and Workshop Cutting-Edge Technologies in Fish and Fisheries Science*, 1, 1 2006.

[28] Michael Cappo, Euan Sinclair Harvey, Hamish A Malcolm, and Peter Speare. Potential of video techniques to monitor diversity, abundance and size of fish in studies of

Marine Protected Areas. 2003. URL https://api.semanticscholar.org/CorpusID:17508271.

[29] Scott A. Chamberlain and Carl Boettiger. `rgbif` for Global Biodiversity Information Facility species occurrence data. *PeerJ Preprints*, 5:e3304v1, 2017. doi: 10.7287/peerj.preprints.3304v1. URL https://doi.org/10.7287/peerj.preprints.3304v1.

[30] Francois Chollet. *Keras*. 2015. URL https://keras.io/. Retrieved April 2022.

[31] Shelley C Clarke, Murdoch K McAllister, E J Milner-Gulland, G P Kirkwood, Catherine G J Michielsens, David J Agnew, Ellen K Pikitch, Hideki Nakano, and Mahmood S Shivji. Global estimates of shark catches using trade records from commercial markets. *Ecology letters*, 9:1115–1126, 10 2006. ISSN 1461-0248 1461-023X. doi: 10.1111/j.1461-0248.2006.00968.x. Place: England.

[32] E Clementi, A Aydogdu, A C Goglio, J Pistoia, R Escudier, M Drudi, A Grandi, A Mariani, V Lyubartsev, R Lecci, S Cretí, G Coppini, S Masina, and N Pinardi. Mediterranean Sea Physical Analysis and Forecast (CMEMS MED-Currents, EAS6 system). *Copernicus Monitoring Environment Marine Service (CMEMS)*, version 1, 2021. doi: 10.25423/cmcc/medsea_analysis_forecast_phy_006_013_eas4.

[33] Andrew P Colefax, Paul A Butcher, and Brendan P Kelaher. The potential for unmanned aerial vehicles (UAVs) to conduct marine fauna surveys in place of manned aircraft. *ICES Journal of Marine Science*, 75:1–8, 6 2017. ISSN 1054-3139. doi: 10.1093/icesjms/fsx100. URL https://doi.org/10.1093/icesjms/fsx100.

[34] Rupert A Collins, Owen S Wangensteen, Eoin J O'Gorman, Stefano Mariani, David W Sims, and Martin J Genner. Persistence of environmental DNA in marine systems. *Communications Biology*, 1:185, 2018. ISSN 2399-3642. doi: 10.1038/s42003-018-0192-6.

[35] LJV Compagno. Relationships of the megamouth shark, Megachasma pelagios (Lamniformes: Megachasmidae), with comments on its feeding habits. *National Oceanic and Atmospheric Administration Technical Report, National Marine Fisheries Service*, 90:357–379, 1990.

[36] Madalyn K Cooper, Roger Huerlimann, Richard C Edmunds, Alyssa M Budd, Agnès Le Port, Peter M Kyne, Dean R Jerry, and Colin A Simpfendorfer. Improved detection sensitivity using an optimal eDNA preservation and extraction workflow and its application to threatened sawfishes. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31:2131–2148, 2021. doi: 10.1002/aqc.3591.

[37] Leanne M Currey-Randall, Mike Cappo, Colin A Simpfendorfer, Naomi F Farabaugh, and Michelle R Heupel. Optimal soak times for Baited Remote Underwater Video Station surveys of reef-associated elasmobranchs. *PLOS ONE*, 15:e0231688–, 5 2020. URL https://doi.org/10.1371/journal.pone.0231688.

[38] K.-F. Dagestad, J Röhrs, Ø Breivik, and B Ådlandsvik. OpenDrift v1.0: a generic framework for trajectory modelling. *Geoscientific Model Development*, 11:1405–1420, 2018. doi: 10.5194/gmd-11-1405-2018.

[39] Jonathan J Dale, Austin M Stankus, Michael S Burns, and Carl G Meyer. The Shark Assemblage at French Frigate Shoals Atoll, Hawai'i: Species Composition, Abundance and Habitat Use. *PLOS ONE*, 6:e16962–, 2 2011. URL https://doi.org/10.1371/journal.pone.0016962.

[40] Simon Dedman, Jerry H Moxley, Yannis P Papastamatiou, Matias Braccini, Jennifer E Caselle, Demian D Chapman, Joshua Eli Cinner, Erin M Dillon, Nicholas K Dulvy, Ruth Elizabeth Dunn, Mario Espinoza, Alastair R Harborne, Euan S Harvey,

Michelle R Heupel, Charlie Huveneers, Nicholas A J Graham, James T Ketchum, Natalie V Klinard, Alison A Kock, Christopher G Lowe, M Aaron MacNeil, Elizabeth M P Madin, Douglas J McCauley, Mark G Meekan, Amelia C Meier, Colin A Simpfendorfer, M Tim Tinker, Megan Winton, Aaron J Wirsing, and Michael R Heithaus. Ecological roles and importance of sharks in the Anthropocene Ocean. *Science*, 385:adl2362, 2024. doi: 10.1126/science.adl2362. URL https://doi.org/10.1126/science.adl2362. doi: 10.1126/science.adl2362 Publisher: American Association for the Advancement of Science.

[41] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12: 2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html. Retrieved January 2022.

[42] Nicholas Dulvy, David Allen, Gina Ralph, and Rachel Walls. *The conservation status of Sharks, Rays and Chimaeras in the Mediterranean Sea*. 12 2016. doi: 10.13140/RG.2.2.22020.53129.

[43] Nicholas K Dulvy, Julia K Baum, Shelley Clarke, Leonard J V Compagno, Enric Cortés, Andrés Domingo, Sonja Fordham, Sarah Fowler, Malcolm P Francis, Claudine Gibson, Jimmy Martínez, John A Musick, Alen Soldo, John D Stevens, and Sarah Valenti. You can swim but you can't hide: the global status and conservation of oceanic pelagic sharks and rays. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 18:459–482, 2008. doi: 10.1002/aqc.975. URL https://doi.org/10.1002/aqc.975.

[44] Nicholas K Dulvy, Sarah L Fowler, John A Musick, Rachel D Cavanagh, Peter M Kyne, Lucy R Harrison, John K Carlson, Lindsay N K Davidson, Sonja V Fordham, Malcolm P Francis, Caroline M Pollock, Colin A Simpfendorfer, George H

Burgess, Kent E Carpenter, Leonard J V Compagno, David A Ebert, Claudine Gibson, Michelle R Heupel, Suzanne R Livingstone, Jonnell C Sanciangco, John D Stevens, Sarah Valenti, and William T White. Extinction risk and conservation of the world's sharks and rays. *eLife*, 3:e00590, 2014. ISSN 2050-084X. URL https://doi.org/10.7554/eLife.00590. Publisher: eLife Sciences Publications, Ltd.

[45] Nicholas K Dulvy, Nathan Pacoureau, Cassandra L Rigby, Riley A Pollom, Rima W Jabado, David A Ebert, Brittany Finucci, Caroline M Pollock, Jessica Cheok, Danielle H Derrick, Katelyn B Herman, C Samantha Sherman, Wade J VanderWright, Julia M Lawson, Rachel H L Walls, John K Carlson, Patricia Charvet, Kinattumkara K Bineesh, Daniel Fernando, Gina M Ralph, Jay H Matsushiba, Craig Hilton-Taylor, Sonja V Fordham, and Colin A Simpfendorfer. Overfishing drives over one-third of all sharks and rays toward a global extinction crisis. *Current Biology*, 31:4773–4787, 2021. ISSN 0960-9822. doi: 10.1016/j.cub.2021.08.062.

[46] Nicholas Dunn, Vincent Savolainen, Sam Weber, Samantha Andrzejaczek, Chris Carbone, and David Curnick. Elasmobranch diversity across a remote coral reef atoll revealed through environmental DNA metabarcoding. *Zoological Journal of the Linnean Society*, 196:593–607, 4 2022. ISSN 0024-4082. doi: 10.1093/zoolinnean/zlac014. URL https://doi.org/10.1093/zoolinnean/zlac014.

[47] Amaya Alvarez 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*, 77:1330–1339, 11 2019. ISSN 1054-3139. doi: 10.1093/icesjms/fsz216.

[48] Bargnesi F, Moro S, Leone A, Giovos I, and Ferretti F. New technologies can support

data collection on endangered shark species in the Mediterranean Sea. *Marine Ecology Progress Series*, 689:57–76, 2022. URL https://www.int-res.com/abstracts/meps/v689/p57-76. 10.3354/meps14030.

[49] Niall Fallon, Sophie Fielding, and Paul Fernandes. Classification of Southern Ocean krill and icefish echoes using random forests. *ICES Journal of Marine Science*, 73: 1998–2008, 2016. ISSN 1054-3139. doi: 10.1093/icesjms/fsw057.

[50] Ian K Fergusson. Distribution and Autecology of the White Shark in the Eastern North Atlantic Ocean and the Mediterranean Sea. pages 321–345, 1996.

[51] L. C. Ferreira and C. Simpfendorfer. *Galeocerdo cuvier*. The IUCN Red List of Threatened Species 2019: e.T39378A2913541. https://dx.doi.org/10.2305/IUCN.UK.2019-1.RLTS.T39378A2913541.en, 2019. Retrieved October 6, 2025.

[52] F Ferretti, J Jenrette, S Moro, C Butner, E Fox, S H D Haddock, S J Jorgensen, T Hastie, and F Micheli. From Data Deficient to Big Data in Shark Conservation. *Fish and Fisheries*, 8 2025. ISSN 1467-2960. doi: https://doi.org/10.1111/faf.70006. URL https://doi.org/10.1111/faf.70006.

[53] Francesco Ferretti, Boris Worm, Gregory L Britten, Michael R Heithaus, and Heike K Lotze. Patterns and ecosystem consequences of shark declines in the ocean. *Ecology Letters*, 13:1055–1071, 8 2010. ISSN 1461-023X. doi: 10.1111/j.1461-0248.2010.01489.x. URL https://doi.org/10.1111/j.1461-0248.2010.01489.x. Publisher: John Wiley & Sons, Ltd.

[54] Francesco Ferretti, Brendan D Shea, Chiara Gambardella, Jeremy F Jenrette, Stefano Moro, Khaled Echwikhi, Robert J Schallert, Austin J Gallagher, Barbara A Block, and Taylor K Chapple. On the tracks of white sharks in the Mediterranean

Sea. *Frontiers in Marine Science*, 11, 2024. ISSN 2296-7745. doi: 10.3389/fmars.2024.1425511. URL https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2024.1425511.

[55] Nathan Fox, Tom August, Francesca Mancini, Katherine E. Parks, Felix Eigenbrod, James M. Bullock, Louis Sutter, and Laura J. Graham. `photosearcher` package in R: An accessible and reproducible method for harvesting large datasets from Flickr. *SoftwareX*, 12:100624, 2020. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2020.100624. URL https://www.sciencedirect.com/science/article/pii/S235271102030337X.

[56] Geoff French, Michal Mackiewicz, Mark Fisher, Helen Holah, Rachel Kilburn, Neil Campbell, and Coby Needle. Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards. *ICES Journal of Marine Science*, 77:1340–1353, 8 2019. ISSN 1054-3139. doi: 10.1093/icesjms/fsz149.

[57] R Froese and D Pauly. Fishbase, 6 2024. URL https://www.fishbase.se/search.php.

[58] C. Gambardella, E. Fernández-Corredor, S. Moro, K. Echwiki, J. F. Jenrette, C. Lemsi, R. J. Schallert, B. D. Shea, M. Chatti Zammit, C. Cerrano, F. Colloca, T. Romeo, J. Navarro, and F. Ferretti. Trophic niche partitioning between the white shark (*Carcharodon carcharias*) and the shortfin mako (*Isurus oxyrinchus*) in the central Mediterranean Sea. *Wildlife Research*, 52:WR25028, 2025. doi: 10.1071/WR25028.

[59] 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*, 77:1354–1366, 10 2019. ISSN 1054-3139. doi: 10.1093/icesjms/fsz186.

[60] Jordan S Goetze, Todd. Bond, Dianne L McLean, Benjamin J Saunders, Tim J Langlois, Steve Lindfield, Laura. A F Fullwood, Damon Driessen, George Shedrawi, and Euan S Harvey. A field and video analysis guide for diver operated stereo-video. *Methods in Ecology and Evolution*, 10:1083–1090, 2019. doi: 10.1111/2041-210X.13189.

[61] Caren S. Goldberg, Cameron R. Turner, Kristy Deiner, Katy E. Klymus, Philip Francis Thomsen, Melanie A. Murphy, Stephen F. Spear, Anna McKee, Sara J. Oyler-McCance, Robert Scott Cornman, Matthew B. Laramie, Andrew R. Mahon, Richard F. Lance, David S. Pilliod, Katherine M. Strickler, Lisette P. Waits, Alexander K. Fremier, Teruhiko Takahara, Jelger E. Herder, and Pierre Taberlet. Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, 7(11):1299–1307, 2016. doi: https://doi.org/10.1111/2041-210X.12595. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12595.

[62] Google Developers. *Google Maps Geocoding API Documentation*. Google LLC, Mountain View, CA, USA, 2025. URL https://developers.google.com/maps/documentation/geocoding/overview.

[63] Chrysoula Gubili, Raşit Bilgin, Evrim Kalkan, S Unsal Karhan, Catherine S Jones, David W Sims, Hakan Kabasakal, Andrew P Martin, and Leslie R Noble. Antipodean white sharks on a Mediterranean walkabout? Historical dispersal leads to genetic discontinuity and an endangered anomalous population. *Proceedings of the Royal Society B: Biological Sciences*, 278:1679–1686, 2011. doi: 10.1098/rspb.2010.1856.

[64] Stephanie E Hampton, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11:

156–162, 4 2013. ISSN 1540-9295. doi: https://doi.org/10.1890/120103. URL https://doi.org/10.1890/120103.

[65] Alexander C Hansell, Steven T Kessel, Lauran R Brewster, Steven X Cadrin, Samuel H Gruber, Gregory B Skomal, and Tristan L Guttridge. Local indicators of abundance and demographics for the coastal shark assemblage of Bimini, Bahamas. *Fisheries Research*, 197:34–44, 2018. ISSN 0165-7836. doi: https://doi.org/10.1016/j.fishres.2017.09.016. URL https://www.sciencedirect.com/science/article/pii/S0165783617302643.

[66] Python hashlib. Encrypting strings using Python hashlib. Python version 3.6, 2021. URL https://docs.python.org/3/library/hashlib.html.

[67] Susanne Hecker, Muki Haklay, anne bowser, Zen Makuch, Johannes Vogel, Aletta Bonn, and Margaret Gold. *Citizen Science – Innovation in Open Science, Society and Policy*. 10 2018. ISBN 978-1-78735-233-9. doi: 10.14324/111.9781787352339.

[68] Michael R Heithaus, Aaron J Wirsing, Lawrence M Dill, and Linda I Heithaus. Long-term movements of tiger sharks satellite-tagged in Shark Bay, Western Australia. *Marine Biology*, 151:1455–1461, 5 2007. ISSN 0025-3162, 1432-1793. doi: 10.1007/s00227-006-0583-y. URL http://link.springer.com/10.1007/s00227-006-0583-y.

[69] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J Belongie. The iNaturalist Challenge 2017 Dataset. *CoRR*, abs/1707.06642, 2017. doi: 10.48550/arXiv.1707.06642.

[70] Gao Huang, Zhuang Liu, and Kilian Q Weinberger. Densely Connected Convolutional Networks. *CoRR*, abs/1608.06993, 2016. doi: 10.48550/arXiv.1608.06993.

[71] IUCN. The IUCN Red List of Threatened Species. *Version 2024-2*, 2024. URL https://www.iucnredlist.org.

[72] J. Jenrette, A. Agustines, E. T. Spencer, R. Schallert, N. Arnoldi, D. Madigan, T. White, K. Koller, J. Berglund, D. Kinzer, B. Block, S. Khalid, and F. Ferretti. Diversifying visual detection and classification artificial intelligence for accessible, semi-automatic monitoring of sharks. .

[73] J. Jenrette, Z. Liu, P. Chimote, T. Hastie, E. Fox, and F. Ferretti. Shark detection and classification with machine learning. *Ecological Informatics*, 69:101673, 2022. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2022.101673. URL https://www.sciencedirect.com/science/article/pii/S1574954122001236.

[74] J. F. Jenrette, J. Jenrette, N. Kobun Truelove, S. Moro, N. Dunn, T. Chapple, A. Gallagher, C. Gambardella, R. Schallert, B. Shea, D. Curnick, B. Block, and F. Ferretti. Detecting Mediterranean White Sharks with Environmental DNA. *Oceanography*, 3 2023. URL https://doi.org/10.5670/oceanog.2023.s1.28.

[75] Jeremy Jenrette, Ariana Agustines, Erin T Spencer, Robert Schallert, Natalie Arnoldi, Daniel Madigan, Tim White, Kelly Koller, Jennifer Berglund, Daniel Kinzer, Barbara Block, Sara Khalid, and Francesco Ferretti. Diversifying visual detection and classification AI for accessible, semi-automatic monitoring of sharks. *Unpublished data*, .

[76] Jeremy Jenrette, Edward Fox, and Francesco Ferretti. Leveraging social networks and open data for ecological observations and population inferences of shark species: Methods, challenges, and opportunities. *Unpublished data*, .

[77] Jeremy Jenrette, Gregory Chang, Steven Gordon, Mason Mulgrew, and Hunter DeBay.

SharkPulse Validation Monitor. *VTechWorks*, 2021. URL fromhttp://hdl.handle.net/10919/103254.

[78] Jeremy Jenrette, Zach Liu, Pranav Chimote, Trevor Hastie, Edward Fox, and Francesco Ferretti. Shark detection and classification with machine learning. *Ecological Informatics*, 69:101673, 2022. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2022.101673. URL https://www.sciencedirect.com/science/article/pii/S1574954122001236.

[79] Jeremy F. Jenrette, S. Khalid, and Francesco Ferretti. Artificial Intelligence for Shark Conservation! Invited presentation (remote) for E/V Nautilus, in partnership with Ocean Exploration Trust and National Geographic, 2022. URL https://www.youtube.com/watch?v=3QZoktX1d0E&list=LL9a7hIAGDsqp--azc5Cp_qw. Presentation.

[80] Jeremy F. Jenrette, Jennifer Jenrette, Nathan Kobun Truelove, Stefano Moro, Nicholas Dunn, Taylor Chapple, Austin Gallagher, Chiara Gambardella, Robert Schallert, Brendan Shea, David Curnick, Barbara Block, and Francesco Ferretti. Detecting Mediterranean White Sharks with Environmental DNA. *Oceanography*, 3 2023. URL https://doi.org/10.5670/oceanog.2023.s1.28.

[81] Yan Jiao, Enric Cortés, Kate Andrews, and Feng Guo. Poor-data and data-poor species stock assessment using a Bayesian hierarchical approach. *Ecological Applications*, 21: 2691–2708, 10 2011. ISSN 1051-0761. doi: https://doi.org/10.1890/10-0526.1. URL https://doi.org/10.1890/10-0526.1.

[82] JSalvador Jorgensen, Fiorenza Micheli, Timothy D White, Kyle S Van Houtan, Joanna Alfaro-Shigueto, Samantha Andrzejaczek, Natalie S Arnoldi, Julia K Baum, Barbara Block, Gregory L Britten, Cheryl Butner, Susana Caballero, Diego Cardeñosa, Taylor

Chapple, Shelley Clarke, Enric Cortes, Nicholas Dulvy, Sarah Louise Fowler, Austin J Gallagher, Eric L Gilman, Brendan Godley, Rachel T Graham, Neil Hammerschlag, Alastair V Harry, Michael Heithaus, Melanie Hutchinson, Charlie Huveneers, Chris G Lowe, Luis O Lucifora, Tracy MacKeracher, Jeffrey Mangel, Ana P Barbosa Martins, Douglas J McCauley, Loren McClenachan, Christopher Mull, Lisa J Natanson, Daniel Pauly, Diana A Pazmiño, Jennifer C A Pistevos, Nuno Queiroz, George Roff, Brendan Shea, Colin Simpfendorfer, David Sims, Christine Ward-Paigeand, Boris Worm, and Francesco Ferretti. Emergent research and priorities for elasmobranch conservation. *Endangered Species Research*, 2022. doi: 10.3354/esr01169.

[83] Julianna P Kadar, Monique A Ladds, Joanna Day, Brianne Lyall, and Culum Brown. Assessment of Machine Learning Models to Identify Port Jackson Shark Behaviours Using Tri-Axial Accelerometers. *Sensors*, 20, 2020. ISSN 1424-8220. doi: 10.3390/s20247096.

[84] Vishnu Kandimalla, Matt Richard, Frank Smith, Jean Quirion, Luis Torgo, and Chris Whidden. Automated Detection, Classification and Counting of Fish in Fish Passages With Deep Learning. *Frontiers in Marine Science*, Volume 8 - 2021, 2022. ISSN 2296-7745. URL https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2021.823173.

[85] Panagiotis Kasapidis and Christina Karli. Optimization of eDNA metabarcoding methodology for the biomontoring of the ichthyofauna in the Eastern Mediterranean Sea. *ARPHA Conference Abstracts*, 4:e65460, 2021. doi: 10.3897/aca.4.e65460.

[86] K Kaschner, K. Kesner-Reyes, C. Garilao, J. Segschneider, T. Rius-Barile J. Rees, and R. Froese. AquaMaps: Predicted range maps for aquatic species, 2019. URL https://www.aquamaps.org.

[87] Brendan P Kelaher, Andrew P Colefax, Alejandro Tagliafico, Melanie J Bishop, Anna Giles, and Paul A Butcher. Assessing variation in assemblages of large marine fauna off ocean beaches using drones. *Marine and Freshwater Research*, 71:68–77, 2020. URL https://doi.org/10.1071/MF18375.

[88] C. Kellogg. DNA Extraction from 0.22µm Sterivex Filters - Phenol-Chloroform. Protocols.io, 2024. URL https://doi.org/10.17504/protocols.io.14egn63oql5d.v1.

[89] Diederik P Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. arXiv, 2014. doi: 10.48550/arXiv.1412.6980.

[90] Aman Kothari, Feneel Patel, Ray Raya, Tirth Shroff, and Ashutosh Tiwari. SharkPulse Validator Game. *VTechWorks*, 2022. URL RetrievedJuly,2023fromhttp://hdl.handle.net/10919/107016.

[91] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F Pereira, C J Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[92] Peter M. Kyne and Rima W. Jabado. Glaucostegus cemiculus. The IUCN Red List of Threatened Species 2019: e.T104050689A104057239, 2019. URL https://dx.doi.org/10.2305/IUCN.UK.2019-2.RLTS.T104050689A104057239.en.

[93] P.M. Kyne, A. Bin Ali, Fahmi, B. Finucci, K. Herman, B.M. Manjaji Matsumoto, and W.J. VanderWright. Chiloscyllium plagiosum. The IUCN Red List of Threatened Species 2021: e.T124554059A124453319, 2021. URL https://dx.doi.org/10.2305/IUCN.UK.2021-1.RLTS.T124554059A124453319.en.

[94] Kevin D Lafferty, Kasey C Benesh, Andrew R Mahon, Christopher L Jerde, and Christopher G Lowe. Detecting Southern California's White Sharks With Environmental DNA. *Frontiers in Marine Science*, 5, 2018. ISSN 2296-7745. doi: 10.3389/fmars.2018.00355.

[95] Yann LeCun and Yoshua Bengio. *Convolutional Networks for Images, Speech, and Time Series*, pages 255–258. MIT Press, 1998. ISBN 0-262-51102-9. URL https://dl.acm.org/doi/10.5555/303568.303704. Retrieved May 2022.

[96] Agostino Leone, Gregory N Puncher, Francesco Ferretti, Emilio Sperone, Sandro Tripepi, Primo Micarelli, Andrea Gambarelli, Maurizio Sarà, Marco Arculeo, Giuliano Doria, Fulvio Garibaldi, Nicola Bressi, Andrea Dall'Asta, Daniela Minelli, Elisabetta Cilli, Stefano Vanni, Fabrizio Serena, Píndaro Díaz-Jaimes, Guy Baele, Alessia Cariani, and Fausto Tinti. Pliocene colonization of the Mediterranean by great white shark inferred from fossil records, historical jaws, phylogeographic and divergence time analyses. *Journal of Biogeography*, 47:1119–1129, 2020. doi: 10.1111/jbi.13794.

[97] Daoliang Li, Qi Wang, Xin Li, Meilin Niu, He Wang, and Chunhong Liu. Recent advances of machine vision technology in fish classification. *ICES Journal of Marine Science*, 79:263–284, 3 2022. ISSN 1054-3139. doi: 10.1093/icesjms/fsab264. URL https://doi.org/10.1093/icesjms/fsab264.

[98] Riverbank Computing Limited. PyQt5 — Comprehensive Python Bindings for Qt v5. https://www.riverbankcomputing.com/software/pyqt/, 2023. Retrieved September 29, 2025.

[99] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in

context. *European conference on computer vision*, 2014. doi: 10.48550/arXiv.1405.0312.

[100] Z Y C Liu, Andy J Chamberlin, Pretom Shome, Isabel J Jones, Gilles Riveau, Raphael A Ndione, Lydie Bandagny, Nicolas Jouanard, Paul Van Eck, Ton Ngo, Susanne H Sokolow, and Giulio A De Leo. Identification of snails and parasites of medical importance via convolutional neural network: an application for human schistosomiasis. *bioRxiv*, 2019. doi: 10.1101/713727.

[101] Ketil Malde, Nils Olav Handegard, Line Eikvil, and Arnt-Børre Salberg. Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77:1274–1285, 4 2019. ISSN 1054-3139. doi: 10.1093/icesjms/fsz057.

[102] Melissa Márquez. Uncovering Mediterranean White Sharks With Environmental DNA. Forbes, April 2023. URL https://www.forbes.com/sites/melissacristinamarquez/2023/04/06/uncovering-mediterranean-white-sharks-with-environmental-dna/. Published April 6, 2023.

[103] Sara Martino, Daniela Silvia Pace, Stefano Moro, Edoardo Casoli, Daniele Ventura, Alessandro Frachea, Margherita Silvestri, Antonella Arcangeli, Giancarlo Giacomini, Giandomenico Ardizzone, and Giovanna Jona Lasinio. Integration of presence-only data from several sources: a case study on dolphins' spatial distribution. *Ecography*, 44:1533–1543, 2021. doi: https://doi.org/10.1111/ecog.05843. URL https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/ecog.05843.

[104] Camilla T McCandless, Bryan S Frazier, James Gelsleichter, and Carolyn N Belcher. Standardized index of abundance for scalloped hammerhead sharks from the NOAA

Fisheries Cooperative Atlantic States Shark Pupping and Nursery longline survey. Technical report, SEDAR, 2021.

[105] Carl Meyer, Jonathan Dale, Yannis Papastamatiou, Nicholas Whitney, and Kim Holland. Seasonal cycles and long-term trends in abundance and species composition of sharks associated with cage diving ecotourism in Hawaii. *Environmental Conservation*, 36:104–111, 6 2009. doi: 10.1017/S0376892909990038.

[106] Carl G Meyer, James M Anderson, Daniel M Coffey, Melanie R Hutchinson, Mark A Royer, and Kim N Holland. Habitat geography around Hawaii's oceanic islands influences tiger shark (Galeocerdo cuvier) spatial behaviour and shark bite risk at ocean recreation sites. *Scientific Reports*, 8:4945, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-23006-0. URL https://doi.org/10.1038/s41598-018-23006-0.

[107] F Migliaccio, D Carrión, and F Ferrario. Semantic Validation of Social Media Geographic Information: A Case Study on Instagram Data for Expo Milano. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4213:1321–1326, 2019. doi: 10.5194/isprs-archives-XLII-2-W13-1321-2019.

[108] Enrico Di Minin, Henrikki Tenkanen, and Tuuli Toivonen. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3: 63, 2015. ISSN 2296-665X. doi: 10.3389/fenvs.2015.00063.

[109] M Miya, Y Sato, T Fukunaga, T Sado, J Y Poulsen, K Sato, T Minamoto, S Yamamoto, H Yamanaka, H Araki, M Kondoh, and W Iwasaki. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science*, 2:150088, 2015. doi: 10.1098/rsos.150088.

[110] Masaki Miya, Tetsuya Sado, Shin ichiro Oka, and Takehiko Fukuchi. The use of citizen science in fish eDNA metabarcoding for evaluating regional biodiversity in a coastal marine region: A pilot study. *Metabarcoding and Metagenomics*, 6:e80444, 2022. doi: 10.3897/mbmg.6.80444. URL https://doi.org/10.3897/mbmg.6.80444.

[111] Stefano Moro, Giovanna Jona-Lasinio, Barbara Block, Fiorenza Micheli, Giulio De Leo, Fabrizio Serena, Massimiliano Bottaro, Umberto Scacco, and Francesco Ferretti. Abundance and distribution of the white shark in the Mediterranean Sea. *Fish and Fisheries*, 21:338–349, 2020. doi: 10.1111/faf.12432.

[112] Stefano Moro, Salvatore Valente, Martina Arcioni, Fabio Falsone, Danilo Scannella, Michele Luca Geraci, Manfredi Di Lorenzo, Giacomo Milisenda, Fabrizio Serena, and Francesco Colloca. Living on the Extinction Edge: Resilience to Fishing and Rebound Potential of the Mediterranean Elasmobranchs. *Fish and Fisheries*, 26(5):772–789, 2025. doi: https://doi.org/10.1111/faf.12911. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/faf.12911.

[113] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814. Omnipress, 2010. ISBN 978-1-60558-907-7. URL https://dl.acm.org/doi/10.5555/3104322.3104425. Retrieved April 2022.

[114] Sridhar Narayan. The generalized sigmoid activation function: Competitive supervised learning. *Information Sciences*, 99:69–82, 1997. ISSN 0020-0255. doi: 10.1016/S0020-0255(96)00200-9.

[115] National Center for Biotechnology Information (NCBI). National Center for Biotechnology Information (NCBI) [Internet]. https://www.ncbi.nlm.nih.gov/, 1988.

Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [cited 2024 Apr 06].

[116] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115:E5716–E5725, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1719367115.

[117] Nathan Pacoureau, Cassie Rigby, Peter Kyne, Richard Sherley, Henning Winker, John Carlson, Sonja Fordham, Rodrigo Barreto, Daniel Fernando, Malcolm Francis, Rima Jabado, Katelyn Herman, Kwang-Ming Lui, Andrea Marshall, Riley Pollom, Evgeny Romanov, Colin Simpfendorfer, Jamie Yin, Holly Kindsvater, and Nicholas Dulvy. Half a century of global decline in oceanic sharks and rays. *Nature*, 589:567–571, 1 2021. doi: 10.1038/s41586-020-03173-9.

[118] Greta Panunzi, Stefano Moro, Isa Marques, Sara Martino, Francesco Colloca, Francesco Ferretti, and Giovanna Jona Lasinio. Estimating the spatial distribution of the white shark in the Mediterranean Sea via an integrated species distribution model accounting for physical barriers. *Environmetrics*, 36(1):e2876, 2025. doi: https://doi.org/10.1002/env.2876. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2876.

[119] E C Pielou. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144, 1966. ISSN 0022-5193. doi: 10.1016/0022-5193(66)90013-0.

[120] R. Pollom, J. Carlson, P. Charvet, C. Avalos, J. Bizzarro, M. P. Blanco-Parra, A. Briones Bell-lloch, M. I. Burgos-Vázquez, D. Cardenosa, A. Cevallos, D. Der-

rick, E. Espinoza, M. Espinoza, P. A. Mejía-Falla, J. M. Morales-Saldaña, A. F. Navia, N. Pacoureau, J. C. Pérez Jiménez, and O. Sosa-Nishizaki. *Sphyrna tiburo* (amended version of 2020 assessment). The IUCN Red List of Threatened Species 2021: e.T39387A205765567. https://dx.doi.org/10.2305/IUCN.UK.2021-3.RLTS.T39387A205765567.en, 2021. Retrieved October 6, 2025.

[121] QIAGEN. *DNeasy Blood & Tissue Kit (Cat. No. 69504)*. QIAGEN, Hilden, Germany, 2024. Used for DNA extraction in laboratory workflows.

[122] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *CoRR*, abs/1804.02767, 2018. doi: 10.48550/arXiv.1804.02767. arXiv: 1804.02767.

[123] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR*, abs/1506.01497, 2015. URL Retrievedfromhttp://arxiv.org/abs/1506.01497.

[124] C. L. Rigby, N. K. Dulvy, R. Barreto, J. Carlson, D. Fernando, S. Fordham, M. P. Francis, K. Herman, R. W. Jabado, K. M. Liu, A. Marshall, N. Pacoureau, E. Romanov, R. B. Sherley, and H. Winker. *Sphyrna lewini*. The IUCN Red List of Threatened Species 2019: e.T39385A2918526. https://dx.doi.org/10.2305/IUCN.UK.2019-1.RLTS.T39385A2918526.en, 2019. Retrieved October 6, 2025.

[125] George Roff, Christopher Doropoulos, Alice Rogers, Yves-Marie Bozec, Nils C Krueck, Eleanor Aurellado, Mark Priest, Chico Birrell, and Peter J Mumby. The Ecological Role of Sharks on Coral Reefs. *Trends in Ecology & Evolution*, 31:395–407, 5 2016. ISSN 0169-5347. doi: 10.1016/j.tree.2016.02.014. URL https://doi.org/10.1016/j.tree.2016.02.014. doi: 10.1016/j.tree.2016.02.014 Publisher: Elsevier.

[126] Armin Ronarcher. Flask, 2024. URL https://palletsprojects.com/.

[127] Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, 346:1063–1064, 2014. ISSN 0036-8075.

[128] Angelique L. Ryan, Cassandra P. O'Hern, and Kelly M. Elkins. Evaluation of Two New Methods for DNA Extraction of "Legal High" Plant Species. *Journal of Forensic Sciences*, 65(5):1704–1708, 2020. doi: https://doi.org/10.1111/1556-4029.14478. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.14478.

[129] Mark Sandler, Andrew G Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. *CoRR*, abs/1801.04381, 2018. URL http://arxiv.org/abs/1801.04381.

[130] Muhammad Saqib, Muhammad Rizwan Khokher, Xin Yuan, Bo Yan, Douglas Bearham, Carlie Devine, Candice Untiedt, Toni Cannard, Kylie Maguire, Geoffrey N Tuck, L Rich Little, and Dadong Wang. Fishing event detection and species classification using computer vision and artificial intelligence for electronic monitoring. *Fisheries Research*, 280:107141, 12 2024. ISSN 0165-7836. doi: 10.1016/j.fishres.2024.107141. URL https://www.sciencedirect.com/science/article/pii/S0165783624002054.

[131] Stefan Schneider, Saul Greenberg, Graham W Taylor, and Stefan C Kremer. Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and Evolution*, 10:3503–3517, 2020. doi: https://doi.org/10.1002/ece3.6147.

[132] Jonas Schroeder. Crawl public Instagram data using R scripts without API access token. *Github*, 2018. URL RetrievedFebruary2021,fromhttps://github.com/JonasSchroeder/InstaCrawlR.

[133] Adam J. Sepulveda, Patrick R. Hutchins, Melissa Forstchen, Megan N. McKeefry, and Amy M. Swigris. The Elephant in the Lab (and Field): Contamination in Aquatic Environmental DNA Studies. *Frontiers in Ecology and Evolution*, 8:609973, 2020. doi: 10.3389/fevo.2020.609973. URL https://doi.org/10.3389/fevo.2020.609973.

[134] F Serena, A J Abella, F Bargnesi, M Barone, F Colloca, F Ferretti, F Fiorentino, J Jenrette, and S Moro. Species diversity, taxonomy and distribution of Chondrichthyes in the Mediterranean and Black Sea. *The European Zoological Journal*, 87:497–536, 2020. doi: 10.1080/24750263.2020.1805518. URL https://www.tandfonline.com/doi/full/10.1080/24750263.2020.1805518. Publisher: Taylor & Francis.

[135] Brendan D Shea, Taylor K Chapple, Khaled Echwikhi, Chiara Gambardella, Jeremy F Jenrette, Stefano Moro, Robert J Schallert, Barbara A Block, and Francesco Ferretti. First satellite track of a juvenile shortfin mako shark (Isurus oxyrinchus) in the Mediterranean Sea. *Frontiers in Marine Science*, 11:1423507, 2024.

[136] C Samantha Sherman, Andrew Chin, Michelle R Heupel, and Colin A Simpfendorfer. Are we underestimating elasmobranch abundances on baited remote underwater video systems (BRUVS) using traditional metrics? *Journal of Experimental Marine Biology and Ecology*, 503:80–85, 6 2018. ISSN 0022-0981. doi: 10.1016/j.jembe.2018.03.002. URL https://www.sciencedirect.com/science/article/pii/S0022098117304781.

[137] C.S. Sherman, C. Simpfendorfer, A. Bin Ali, D. Derrick, Dharmadi, Fahmi, D. Fernando, A.B. Haque, A. Maung, L. Seyha, D. Tanay, J.A.T. Utzurrum, V.Q. Vo, and R.R. Yuneni. Taeniura lymma. The IUCN Red List of Threatened Species 2021: e.T116850766A116851089, 2021. URL https://dx.doi.org/10.2305/IUCN.UK.2021-1.RLTS.T116850766A116851089.en.

[138] Arial J Shogren, Jennifer L Tank, Elizabeth Andruszkiewicz, Brett Olds, Andrew R Mahon, Christopher L Jerde, and Diogo Bolster. Controls on eDNA movement in streams: Transport, Retention, and Resuspension. *Scientific Reports*, 7:5065, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-05223-1. URL https://doi.org/10.1038/s41598-017-05223-1.

[139] Shoaib Ahmed Siddiqui, Ahmad Salman, Muhammad Imran Malik, Faisal Shafait, Ajmal Mian, Mark R Shortis, and Euan S Harvey. Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. *ICES Journal of Marine Science*, 75:374–389, 7 2018. ISSN 1054-3139. doi: 10.1093/icesjms/fsx109.

[140] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 2015. doi: 10.48550/arXiv.1409.1556.

[141] C. Simpfendorfer, Fahmi, A. Bin Ali, D., J. A. T. Utzurrum, L. Seyha, A. Maung, K. K. Bineesh, R. R. Yuneni, A. Sianipar, A. B. Haque, D. Tanay, D. A. Gautama, and V. Q. Vo. *Carcharhinus amblyrhynchos*. The IUCN Red List of Threatened Species 2020: e.T39365A173433550. https://dx.doi.org/10.2305/IUCN.UK.2020-3.RLTS.T39365A173433550.en, 2020. Retrieved October 6, 2025.

[142] Tao Song, Cong Pang, Boyang Hou, Guangxu Xu, Junyu Xue, Handan Sun, and Fan Meng. A review of artificial intelligence in marine science. *Frontiers in Earth Science*, 11, 2023. ISSN 2296-6463. doi: 10.3389/feart.2023.1090185. URL https://www.frontiersin.org/articles/10.3389/feart.2023.1090185.

[143] Felicia Spencer. eDNA testing for sharks in the Mediterranean Sea yields fin-tastic results. Virginia Tech News, July 2025. URL https://news.vt.edu/articles/2025/

07/white-sharks-in-mediterranean-flsi-jenrette-genetic-cnre-.html. Published July 21, 2025.

[144] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. URL RetrievedFebruary2022,fromhttp://jmlr.org/papers/v15/srivastava14a.html.

[145] Phillip C Sternes and Kenshu Shimada. Body forms in sharks (Chondrichthyes: Elasmobranchii) and their functional, ecological, and evolutionary implications. *Zoology*, 140:125799, 2020. ISSN 0944-2006. doi: https://doi.org/10.1016/j.zool.2020.125799. URL https://www.sciencedirect.com/science/article/pii/S0944200620300581.

[146] JD Stevens, Ramon Bonfil, Nicholas K Dulvy, and PA Walker. The effects of fishing on sharks, rays, and chimaeras (chondrichthyans), and the implications for marine ecosystems. *ICES Journal of Marine Science*, 57(3):476–494, 2000.

[147] Tiziano Storai, Antonio Celona, Marco Zuffa, and Alessandro De Maddalena. On the occurrence of the porbeagle, Lamna nasus (Bonnaterre, 1788) (Chondrichthyes: Lamnidae), off Italian coasts (northern and central Mediterranean Sea): A historical survey. In *Annales: Series Historia Naturalis*, volume 15, page 195. Scientific and Research Center of the Republic of Slovenia, 2005.

[148] James Sulikowski, Carolyn Wheeler, Austin Gallagher, Bianca Prohaska, Joseph Langan, and Neil Hammerschlag. Seasonal and life-stage variation in the reproductive ecology of a marine apex predator, the tiger shark Galeocerdo cuvier, at a protected female-dominated site. *Aquatic Biology*, 24, 2 2016. doi: 10.3354/ab00648.

[149] Brian L Sullivan, Jocelyn L Aycrigg, Jessie H Barry, Rick E Bonney, Nicholas Bruns, Caren B Cooper, Theo Damoulas, André A Dhondt, Tom Dietterich, Andrew Farnsworth, Daniel Fink, John W Fitzpatrick, Thomas Fredericks, Jeff Gerbracht, Carla Gomes, Wesley M Hochachka, Marshall J Iliff, Carl Lagoze, Frank A La Sorte, Matthew Merrifield, Will Morris, Tina B Phillips, Mark Reynolds, Amanda D Rodewald, Kenneth V Rosenberg, Nancy M Trautmann, Andrea Wiggins, David W Winkler, Weng-Keen Wong, Christopher L Wood, Jun Yu, and Steve Kelling. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014. ISSN 0006-3207. doi: 10.1016/j.biocon.2013.11.003.

[150] Mark Sullivan, Stacie Robinson, and Charles Littnan. Social media as a data resource for monkseal conservation. *PLOS ONE*, 14:1–11, 10 2019. doi: 10.1371/journal.pone.0222627. URL https://doi.org/10.1371/journal.pone.0222627. Publisher: Public Library of Science.

[151] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, 2:150026, 6 2015. ISSN 2052-4463. doi: 10.1038/sdata.2015.26.

[152] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C Vercauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, Ben Teton, James C Beasley, Peter E Schlichting, Raoul K Boughton, Bethany Wight, Eric S Newkirk, Jacob S Ivan, Eric A Odell, Ryan K Brook, Paul M Lukacs, Anna K Moeller, Elizabeth G Mandeville, Jeff Clune, and Ryan S Miller. Machine learning to classify animal species in camera trap images:

Applications in ecology. *Methods in Ecology and Evolution*, 10:585–590, 2019. doi: 10.1111/2041-210X.13120.

[153] Chris Taklis, Ioannis Giovos, and Alexandros Karamanlidis. Social media: a valuable tool to inform shark conservation in Greece. *Mediterranean Marine Science*, 6 2020. doi: 10.12681/mms.22165.

[154] Luke Taylor and Geoff Nitschke. Improving Deep Learning using Generic Data Augmentation. *CoRR*, abs/1708.06020, 2017. doi: 10.48550/arXiv.1708.06020.

[155] the Apify team Jan Čurn Jakub Balada. Apify: Web Scraping Solutions, 2015. URL https://apify.com/.

[156] Mikhail Fursov the UGENE team Konstantin Okonechnikov Olga Golosova. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, 28:1166–1167, 2012. doi: doi:10.1093/bioinformatics/bts091.

[157] A. C. Thomas, J. Howard, P. L. Nguyen, T. A. Seimon, and C. S. Goldberg. eDNA Sampler: A fully integrated environmental DNA sampling system. *Methods in Ecology and Evolution*, 9:1379–1385, 2018. doi: 10.1111/2041-210X.12994.

[158] Philip Francis Thomsen, Peter Rask Møller, Eva Egelyng Sigsgaard, Steen Wilhelm Knudsen, Ole Ankjær Jørgensen, and Eske Willerslev. Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PloS one*, 11(11):e0165252, 2016.

[159] Tuuli Toivonen, Vuokko Heikinheimo, Christoph Fink, Anna Hausmann, Tuomo Hiippala, Olle Järv, Henrikki Tenkanen, and Enrico Di Minin. Social media data for conservation science: A methodological overview. *Biological Conservation*, 233:298 – 315, 2019. ISSN 0006-3207. doi: 10.1016/j.biocon.2019.01.023.

[160] Julien Troudet, Philippe Grandcolas, Amandine Blin, Régine Vignes-Lebbe, and Frédéric Legendre. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, 7(1):9132, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-09084-6. URL https://doi.org/10.1038/s41598-017-09084-6.

[161] Nathan K Truelove, Elizabeth A Andruszkiewicz, and Barbara A Block. A rapid environmental DNA method for detecting white sharks in the open ocean. *Methods in Ecology and Evolution*, 10:1128–1135, 2019. doi: 10.1111/2041-210X.13201.

[162] Aloysius T M van Helmond, Lars O Mortensen, Kristian S Plet-Hansen, Clara Ulrich, Coby L Needle, Daniel Oesterwind, Lotte Kindt-Larsen, Thomas Catchpole, Stephen Mangi, Christopher Zimmermann, Hans Jakob Olesen, Nick Bailey, Heidrikur Bergsson, Jørgen Dalskov, Jon Elson, Malo Hosken, Lisa Peterson, Howard McElderry, Jon Ruiz, Johanna P Pierre, Claude Dykstra, and Jan Jaap Poos. Electronic monitoring in fisheries: Lessons from global experiences and future opportunities. *Fish and Fisheries*, 21:162–189, 2020. doi: 10.1111/faf.12425. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/faf.12425.

[163] Anthony van Rooyen, Adam D Miller, Zach Clark, Craig D H Sherman, Paul A Butcher, Justin R Rizzari, and Andrew R Weeks. Development of an environmental DNA assay for detecting multiple shark species involved in human–shark conflicts in Australia. *Environmental DNA*, 3:940–949, 2021. doi: 10.1002/edn3.202.

[164] Filippo Varini, Jeremy Jenrette, Francesco Ferretti, Joel Gayford, Mark Bond, Matthew Witt, Sophie Wilday, and Ben Glocker. SharkTrack: an accurate, generalisable software for streamlining shark and ray underwater video analysis. *Accepted in Ecological Informatics*, 2025.

[165] Adriana Vella and Joseph G Vella. Central-southern Mediterranean submarine canyons

and steep slopes: role played in the distribution of cetaceans, bluefin tunas and elasmobranchs. 2012. URL https://www.um.edu.mt/library/oar/handle/123456789/93408. Publisher: International Union for Conservation of Nature and Natural Resources. Retrieved Oct 15, 2022.

[166] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, 41:24–32, 2017. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2017.07.004. URL https://www.sciencedirect.com/science/article/pii/S1574954116302047.

[167] Sébastien Villon, David Mouillot, Marc Chaumont, Emily S Darling, Gérard Subsol, Thomas Claverie, and Sébastien Villéger. A Deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecological Informatics*, 48:238–244, 2018. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2018.09.007. URL https://www.sciencedirect.com/science/article/pii/S1574954118300694.

[168] Sébastien Villon, Corina Iovan, Morgan Mangeas, and Laurent Vigliola. Toward an artificial intelligence-assisted counting of sharks on baited video. *Ecological Informatics*, 80:102499, 5 2024. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2024.102499. URL https://www.sciencedirect.com/science/article/pii/S1574954124000414.

[169] Ben G Weinstein. Scene-specific convolutional neural networks for video-based biodiversity detection. *Methods in Ecology and Evolution*, 9:1435–1441, 2018. doi: 10.1111/2041-210X.13011.

[170] W T White and P R Last. A review of the taxonomy of chondrichthyan fishes: a modern perspective. *Journal of Fish Biology*, 80:901–917, 4 2012. ISSN 0022-1112. doi:

10.1111/j.1095-8649.2011.03192.x. URL https://doi.org/10.1111/j.1095-8649.2011.03192.x. Publisher: John Wiley & Sons, Ltd.

[171] Robin C Whytock, Jędrzej Świeżewski, Joeri A Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa el din, Kelly Boekee, Stephanie Brittain, Anabelle W Cardoso, Philipp Henschel, David Lehmann, Brice Momboua, Cisquet Kiebou Opepa, Christopher Orbell, Ross T Pitman, Hugh S Robinson, and Katharine A Abernethy. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, 12:1080–1092, 2021. doi: 10.1111/2041-210X.13576.

[172] John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Tim Robertson, David Vieglais, Chris Spencer, Kirk Schulz, Walter Döring, et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, 7(1):e29715, 2012. doi: 10.1371/journal.pone.0029715. URL https://doi.org/10.1371/journal.pone.0029715.

[173] Erik S. Wright. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal*, 8(1):352–359, 2016.

[174] Chi-Ju Yu, Shoou-Jeng Joung, Hua-Hsun Hsu, Chia-Yen Lin, Tzu-Chi Hsieh, Kwang-Ming Liu, and Atsuko Yamaguchi. Spatial–Temporal Distribution of Megamouth Shark, Megachasma pelagios, Inferred from over 250 Individuals Recorded in the Three Oceans. *Animals*, 11, 2021. ISSN 2076-2615. doi: 10.3390/ani11102947. URL https://www.mdpi.com/2076-2615/11/10/2947.

[175] Hongkun Yu, Chen Chen, Xianzhi Du, Yeqing Li, Abdullah Rashwan, Le Hou, Pengchong Jin, Fan Yang, Frederick Liu, Jaeyoun Kim, and Jing Li. TensorFlow Model Garden, 2020. URL https://github.com/tensorflow/models.

[176] J Yuan, C Chen, W Yang, M Liu, J Xia, and S Liu. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7:3–36, 2021. ISSN 2096-0662. doi: 10.1007/s41095-020-0191-7.

# Appendices

# Appendix A

# Software and Repositories

This appendix provides a list of all public repositories, web resources, and software developed or co-developed during this dissertation. These resources are openly available to facilitate continued development, reproducibility, and collaboration. As allowed by publication or requested collaboration, more repositories may become available. Researchers and developers are encouraged to explore, extend, or integrate these tools to advance the next generation of digital and molecular wildlife monitoring.

- **The Shark Detector (SD)** — github.com/sharkPulse/sharkDetectoR
  R package and framework for automated shark detection, filtering, and species-level classification in visual media.
- **SD API and Developer's Repository** — github.com/sharkPulse/sharkdetector-dev
  Flask-based API and developer environment supporting the Shark Detector's back-end operations and model deployment.
- **SharkByte Application** — github.com/sharkPulse/sdapp
  Cross-platform desktop and mobile graphical interface for running Shark Detector models locally and viewing annotated results.
- **SharkByte Instructions and Documentation** — sp2.cs.vt.edu/applications/sharkbyte
  Online guide for installing, running, and troubleshooting the SharkByte application, including example workflows.

- **Social Network Modeling Repository** — github.com/sharkPulse/sp-sn
  R and Python scripts for sourcing, cleaning, and modeling shark observations from social networks and online archives.

- **Particle Tracking with OpenDrift** — github.com/JeremyFJ/particle-abundance
  Python-based workflow for oceanographic particle modeling and visualization used to simulate Environmental DNA (eDNA) transport and degradation dynamics.

- **Other Public SharkPulse Repositories** — github.com/sharkPulse
  Central hub for additional open-source tools, data pipelines, and utilities developed under the SharkPulse initiative.

Together, these resources provide an open and extensible foundation for continuing the monitoring initiatives established in this dissertation. They are designed for transparency, collaboration, and practical implementation in future research and management applications.

# Appendix B

# Instagram dataset

From August to December 2021, we assembled an Instagram (IG) dataset by scraping the most recent posts from 14 shark-related hashtags (e.g., `#sandtigershark`). The raw posts were first filtered with the Shark Identifier (SI) to distinguish shark from non-shark content, after which the Shark Classifier (SC) was applied to classify images to the species level.

In the initial implementation of the Shark Detector (SD)v1, the SI achieved 91% accuracy for binary classification of shark versus non-shark posts (Figure B.1), while the SC reached 69% top recall and 76% top-3 recall for species-level predictions (Figure B.2). With the expanded dataset and hierarchical framework in SDv5, the SI maintained comparable performance (91%), but the SC improved substantially, achieving 80% top recall and 90% top-3 recall across 19 species, resulting in 3,036 unique classified observations. These results highlight the improvement in species-level performance between versions and demonstrate the feasibility of using social media data as a source of high-resolution occurrence records.

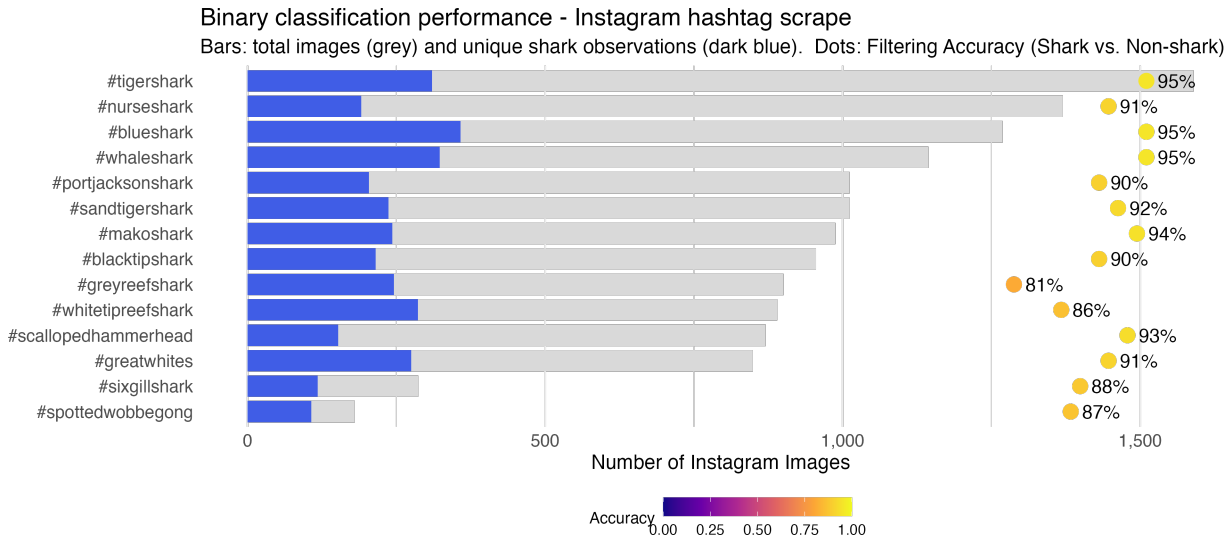


Figure B.1: Results of classifying 14 IG hashtags using the SI. The gray bars indicate the total amount of retrieved images, while the blue bars represent shark images. The SI classification performance is indicated as colored dots on the right.

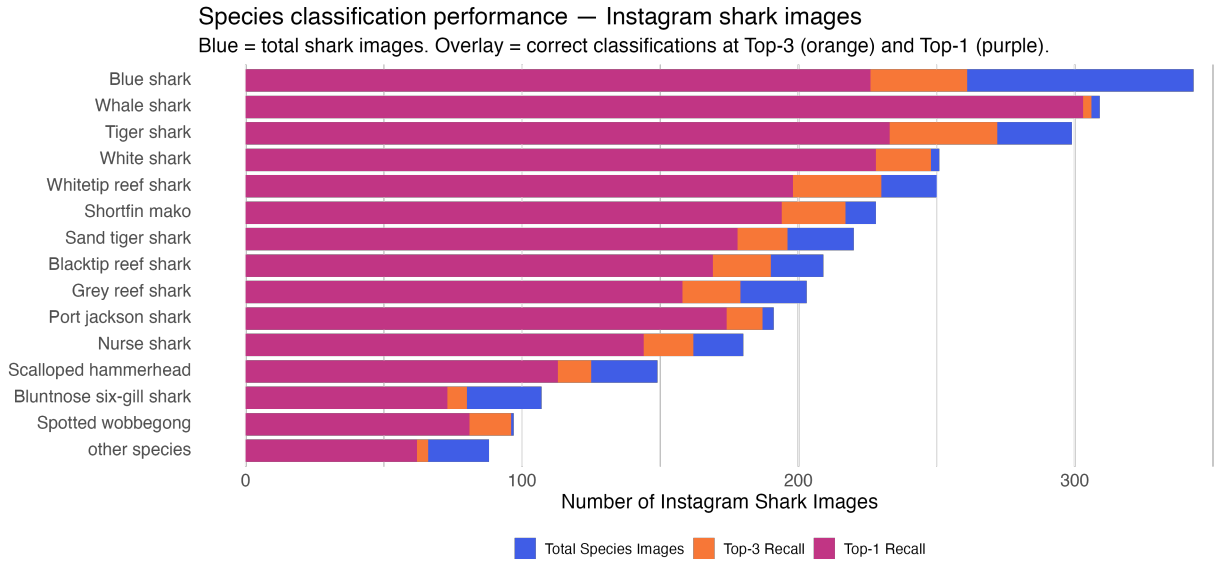


Figure B.2: Results of classifying 19 shark species with the SC. Colored bars represent the total amount of shark images (blue), the amount of images correctly classified with three guesses (orange), and the amount of images classified with one guess (maroon).

# Appendix C

# Citizen Science and Model Augmentation

This appendix outlines how citizen scientists and SharkByte users can contribute annotated shark images to sharkPulse, thereby strengthening the training datasets and improving future versions of the Shark Detector. This workflow not only improves the classification performance of the SD, strengthening sharkPulse, but also ensures more accurate, boosted versions of the publicly available `sharkDetectoR` and SharkByte software for future iterations. We provide an instructive workflow that highlights shark species most likely to occur within a given region, guiding users toward relevant video material. These videos can then be processed with SharkByte and submitted to sharkPulse, boosting training datasets and improving classification performance for regionally important species. To demonstrate this workflow and its results, we tested it on the Main Hawaiian Island (MHI) and Palau surveys, with the objective of boosting species-level classification for all sharks observed in both regions (Figure C.1).

## C.1 Identifying Regional Species

First, we supplied geographic bounding boxes to the `find_species` function of the `sharkDetectoR` package for both survey regions and generated a list of shark species.

For supplying precise bounding box coordinates, we chose to calculate an approximate 80-kilometer radius from the Palauan archipelago's centroid and 270 kilometers from the Hawaiian archipelago's centroid, then convert to degrees of latitude and longitude for a spatial square. We then filtered the list, retaining only shark species that are classifiable by the SC and contain $> 50\%$ likelihood to be observed in either region (determined by averaging the bounded spatial AquaMaps prediction). This approach resulted in 13 shark species in Palau and 17 species in the Hawaiian archipelago with a known probability of occurrence (Table C.1).

To collect visual training data, we then sourced short YouTube (YT) videos of the species. We were unable to include four species (*Nebrius ferrugineus*, *Megachasma pelagios*, *Isurus paucus*, and *Alopias superciliosus*) due to their lower encounter rates.

We then processed each video with SharkByte, producing whole frames, cropped images, and an annotation spreadsheet. Each video was processed at a fixed detection threshold of 0.5, which favors cleaner data even if some sharks are occasionally missed. Videos were analyzed at 15 frames per second, and we prioritized short clips ($< 3$ minutes) containing only one species to reduce manual validation effort. SharkByte produced full frames, cropped images, and an annotation spreadsheet. Using the built-in metadata fields, we appended species identity, a general location name, centroid coordinates, video source, and relevant notes to the annotation file.

## C.2 Submitting Data to sharkPulse

After manually validating the processed 16 videos describing 16 species, we compressed the output using the SharkByte button "Zip Output Folder." Then, we registered a set of credentials on the sharkPulse website (`https://sharkpulse.org/applications/sharkbyte/`)

with an email, a username, and a secure password. The registration webpage provides clear instructions on setting up valid credentials, which helps sharkPulse manage data submissions, prevent spam or malicious use, track individual contributions, and incentivize continued user engagement through specialized events and rewards. Each individual submission is limited to 1 GB to prevent server overloading and streamline manual review. The user is encouraged to submit both full-frame images of detected sharks, which capture essential background context and environmental complexity, as well as cropped images that isolate the subjects, ensuring the training dataset includes examples with and without surrounding visual noise. We then submitted the compressed validated images to sharkPulse using the "Upload Zip to Server" SharkByte button, making sure to provide the correct registered credentials. When connected to the internet, the data is uploaded to sharkPulse, where it is decompressed, and the annotations and media are securely stored for expert review.

The submitted data were reviewed by the sharkPulse expert who demonstrated this workflow (Jeremy Jenrette). The annotated images were then piped into the SC training dataset, prompting retraining. New video frames were added to the training and validation sets (at a 9:1 split), leaving the holdout test set unchanged. This approach isolated the impact of training data augmentation on model performance, enabling an unbiased evaluation of improvements in classification recall. We tested the retrained SC on the same 46,332 sharkPulse holdout test images, and the subset of footage from both surveys. To further promote this application and workflow for citizen scientists—by identifying relationships between manual effort and direct changes to model performance—we plotted the number of new images boosted per species versus the corresponding change in species-specific classification recall for the test images and survey subsets (Figure C.1).

## C.3 Data Augmentation

In less than three hours, we sourced a list of 24 species likely to be observed in Palau or the MHIs, collected 20 videos (one per species, excluding four cryptic species) from YT with an average duration of 73 ± 57 seconds, processed them with SharkByte, and submitted the output to the sharkPulse Application Programming Interface (API) for expert review (Figure C.1). All species that were observed in both surveys were present in the final list and represented with new training data. We sourced location metadata for 16 videos and produced 8,466 new images comprised of whole detected frames and cropped shark subjects.

It took 8.5 hours to retrain the SC and 16 minutes to test it on the same 46,332 holdout sharkPulse images with the `sharkDetectoR`. Species-specific recall improved by an average of 7.7% ± 1.3% following targeted boosting, while overall end-to-end accuracy rose by only 0.6% to 92.0%, due to minor reductions in recall for non-boosted species because of increased misclassification into boosted classes. However, this trade-off was less consequential in practical survey applications, as non-target species were not observed in the survey videos, resulting in a more pronounced improvement of 9.5% ± 3.5% in survey recall.

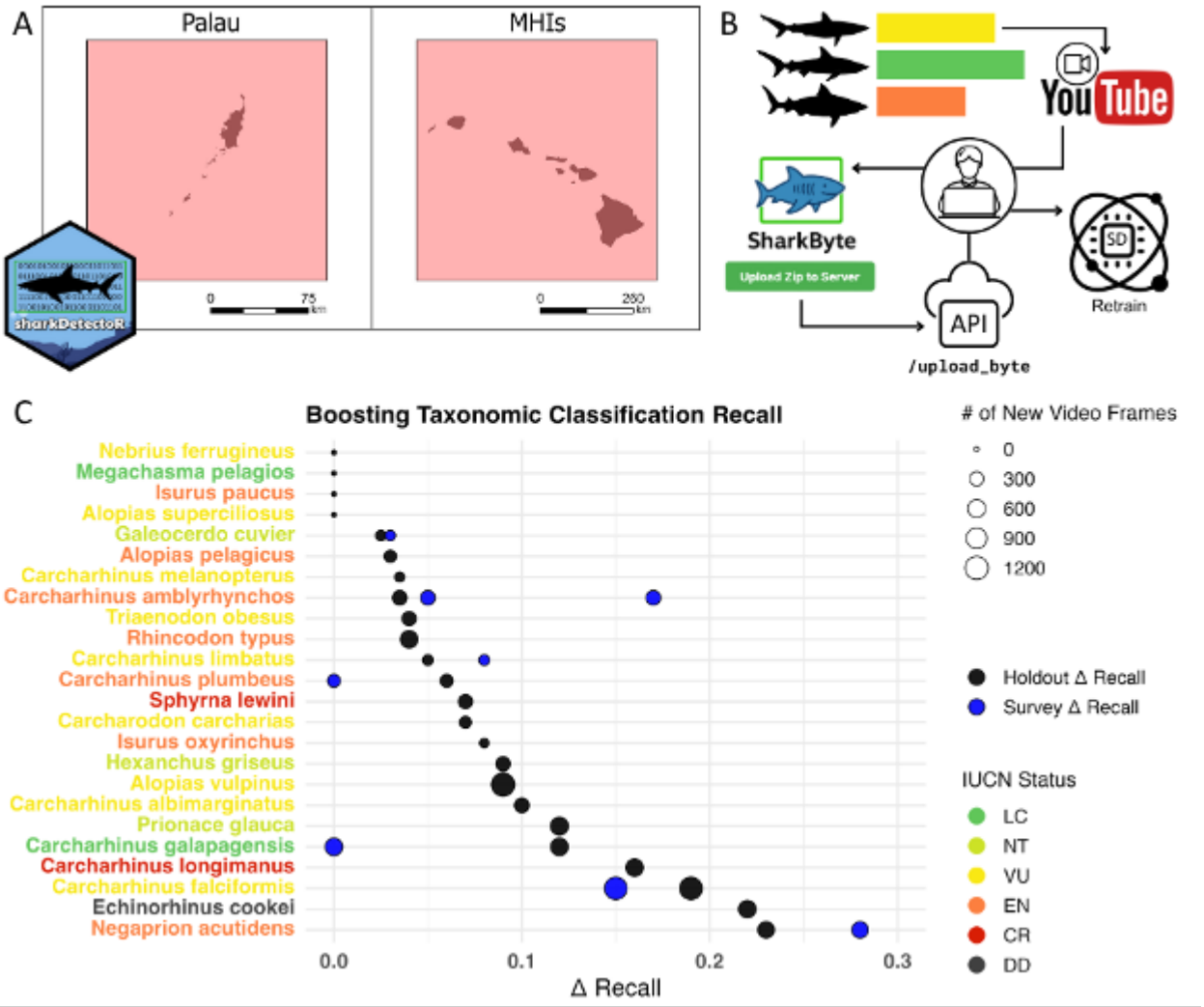


Figure C.1: Boosting base and survey-specific recall. The data augmentation workflow identifies shark species in a geographic bound, processes relevant video training data, and submits data to sharkPulse for updating species classification performance. In panel (A) we used the `sharkDetectoR::find_species` function to identify 20 unique species with probable residency in Hawaii and Palau. In (B), the list of shark species guided manually sourcing relevant video footage from YT and processing them with the SharkByte application. Processed media was submitted to sharkPulse and retrained. In (C), we evaluated the result of this approach to increase base and survey-specific classification performance. Species are colored by their International Union for Conservation of Nature (IUCN) conservation status, Δ Recall points are sized by how many new images were trained, and performance was measured on the holdout test and survey datasets.

Table C.1: Catalogued information on shark species in Palau and the MHIs, integrating IUCN distributions, FishBase depth ranges, and AquaMaps occurrence probabilities. Also shown are the number of images sourced from YT videos via SharkByte and the resulting base recall increase of the SD after retraining.

| Species | PA prob | MHI prob | Category | Max depth (m) | Min depth (m) | Images | Recall Increase |
|---|---|---|---|---|---|---|---|
| *Alopias pelagicus* | 0.64 | – | EN | 300 | 0 | 1,327 | 0.07 |
| *Alopias superciliosus* | 0.82 | 0.98 | VU | 730 | 0 | 1,821 | 0.04 |
| *Carcharhinus albimarginatus* | 0.89 | 0.84 | VU | 800 | 0 | 594 | 0.01 |
| *Carcharhinus amblyrhynchos* | 0.90 | 0.85 | EN | 1000 | 0 | 2,100 | 0.10 |
| *Carcharhinus falciformis* | 0.55 | – | VU | 4000 | 0 | 4,893 | 0.13 |
| *Carcharhinus longimanus* | 0.65 | 0.62 | CR | 1082 | 0 | 6,540 | 0.03 |
| *Carcharhinus melanopterus* | 1.00 | 0.87 | VU | 75 | 0 | 1,518 | 0.03 |
| *Isurus oxyrinchus* | 0.75 | 1.00 | EN | 888 | 0 | 1,159 | 0.09 |
| *Isurus paucus* | 0.79 | 1.00 | EN | 1752 | 0 | 722 | 0.25 |
| *Negaprion acutidens* | 0.95 | – | EN | 92 | 0 | 844 | 0.14 |
| *Prionace glauca* | 0.74 | 1.00 | NT | 1082 | 0 | 8,104 | 0.17 |
| *Rhincodon typus* | 0.93 | 1.00 | EN | 1928 | 0 | 491 | 0.01 |
| *Triaenodon obesus* | 0.98 | 0.97 | VU | 330 | 0 | 8,900 | 0.14 |
| *Alopias vulpinus* | 0.66 | – | VU | 650 | 0 | 590 | 0.02 |
| *Carcharhinus galapagensis* | 1.00 | – | LC | 286 | 1 | 722 | 0.11 |
| *Carcharhinus limbatus* | 0.90 | – | VU | 140 | 0 | 610 | 0.07 |
| *Carcharhinus plumbeus* | 0.99 | – | EN | 500 | 0 | 6,793 | 0.09 |
| *Carcharodon carcharias* | 0.96 | – | VU | 1280 | 0 | 9,190 | 0.15 |
| *Galeocerdo cuvier* | 0.86 | – | NT | 800 | 0 | 7,100 | 0.19 |
| *Sphyrna lewini* | 0.67 | – | CR | 1043 | 0 | 1,499 | 0.11 |

### C.3.1 Impact

Visual observations describing some species, such as *Megachasma pelagios* (megamouth shark), with approximately 250-300 total documented observations in the wild [174], cannot be substantially boosted without synthetic data-generation techniques [2] or similar methods, exposing a current limitation but also a future step towards resolving an imbalanced training dataset. However, results showed that even data-deficient species such as *Echinorhinus cookei* (prickly shark) could be sourced and boosted with 1,100 new images producing a recall increase of 22% (Figure C.1). Despite significant gains in species-specific recall, overall end-to-end accuracy improved modestly. This phenomenon underscores the complex trade-offs inherent in targeted data augmentation strategies within multi-class classification systems [78, 130]. By using the `get_metrics` function to reveal the number of training images per species, this approach can also prioritize data-deficient species to balance the dataset

and directly tackle key limitations in species-specific accuracy [78, 84]. Therefore, the data augmentation strategy provided a substantial survey-specific performance boost, exemplified by a 9.5% increase in taxonomic recall, which emphasizes the practical benefits of selectively improving species-specific model performance. To further capitalize on newly ingested data, next steps should incorporate automatic boosting of the object detection model to fully augment all prediction outputs.

The `sharkDetectoR` and SharkByte tools allowed rapid, streamlined sourcing of shark distribution probabilities for guiding data collection and annotation of shark videos from platforms such as YT. This approach proved effective for both programmatic and practical field use by generating 8,466 new training data (6,959 spatially tagged) that boosted the base recall of the SC and survey-specific performance by 0.6% and 9.5% respectively. While this method does not yet automatically source video data given a list of target species, it provides a structured scheme to direct manual effort toward the most ecologically and conservation-relevant taxa, thereby improving the detection capability and accuracy of automated postprocessing. Crucially, this framework demonstrates how global, publicly available media streams can be converted into validated biodiversity records that augment conventional datasets. By prioritizing species of conservation concern, incorporating occurrence likelihood, and integrating metadata from focal regions, the workflow not only improves classifier accuracy but also produces actionable ecological information. In practical terms, this means managers and conservation organizations can access near-real-time updates on species presence, track shifts in local shark communities, and identify emerging hotspots of biodiversity or risk. More broadly, the study illustrates how combining Artificial Intelligence (AI) with citizen-generated data can reduce barriers to large-scale monitoring, enabling more responsive and evidence-based strategies for shark conservation and management.

# Appendix D

# Morphology and taxonomy

This appendix provides detailed summaries of the SD's taxonomic coverage and morphological analyses, including modeling how species' morphometric distinctness and data availability influence classification performance across orders, families, genera, and species.

## D.1 Predicting Performance with Morphometrics

Automated classification of sharks presents unique challenges due to the wide range of morphological similarities across closely related species. Many elasmobranch taxa are defined by subtle differences in body shape, fin proportions, and coloration. These are features that can be difficult to consistently distinguish in photographs or video frames. These difficulties are compounded when images vary in quality, perspective, or background complexity, all of which contribute noise to the classification task. For Convolutional Neural Networks (CNNs), such morphological overlap can produce systematic misclassifications, particularly when two or more species exhibit similar morphometric measurements. In these cases, the predictive capacity of a model is constrained not only by the number of available training images but also by the inherent distinctness of the taxa themselves. Thus, assessing the degree to which species are morphometrically separable is critical for understanding where classification bottlenecks arise and for guiding targeted data augmentation efforts.

To address this issue, we modeled the predictive relationship between classification per-

formance and the morphological distinctness of shark taxa. Morphometric data were compiled from the `rfishbase` package [17], which aggregates curated measurements across body length, fin ratios, and other diagnostic characters. For each taxonomic rank (order, family, genus, and species), we calculated multivariate centroids that capture the average morphometric profile of each group. Pairwise Euclidean distances between these centroids were then used to quantify morphological distinctness at each taxonomic level. Classification performance was summarized using the joint $F_1$ Score ($F_1$) score from the SD pipeline, a metric reflecting both precision and recall. To account for imbalances in training data availability, we incorporated the total number of training images per species as weights in the regression analysis. This framework allowed us to directly test whether greater morphological distinctness predicts improved classification outcomes, while controlling for training data volume.

Our analysis included 47 species, 22 genera, 13 families, and five orders. A weighted linear regression revealed that morphological distinctness was a significant predictor of classification performance (Figure D.1). Formally, we modeled the relationship as:

$$F_{1,i} = \beta_0 + \beta_1 \cdot \log(D_i) + \varepsilon_i, \tag{D.1}$$

where $F_{1,i}$ is the joint classification performance of taxon $i$, $D_i$ is the morphometric distance of that taxon from the centroid of its training dataset, $\beta_0$ and $\beta_1$ are estimated regression coefficients, and $\varepsilon_i$ is the error term. Model weights were assigned according to the number of training images available per species, ensuring that more data-rich taxa had proportionally greater influence on the regression fit. The fitted model indicated that the log of taxa-centroid distance was positively associated with joint $F_1$ scores ($\beta = 0.035$, $p = 0.031$). The intercept term (0.783) reflected baseline performance even for morphologically

similar species, whereas increases in morphological distance were associated with incremental gains in classification accuracy. Despite a relatively low proportion of variance explained ($R^2 = 0.099$), the relationship was statistically significant, underscoring that morphological distinctness provides useful predictive power when combined with training data availability. Residual variability, however, highlights that other factors—such as image heterogeneity, habitat context, and labeling noise—also influence performance.

These findings have several implications for strategically improving the SD. First, taxa with low morphological distinctness represent high-priority targets for data augmentation, as they are more likely to be misclassified without substantial additional training material. Conversely, morphologically distinct species can often achieve high performance with fewer images, suggesting that augmentation efforts there may yield diminishing returns. More broadly, incorporating morphometric predictors into performance modeling provides a framework for balancing training datasets and anticipating classification challenges. This approach can be generalized across taxonomic levels, providing researchers with a diagnostic tool for identifying species most likely to benefit from targeted boosting. Ultimately, by combining crowdsourced morphometric data with training image curation, we can better align machine learning workflows with biological reality, yielding more robust and ecologically meaningful classification outcomes.

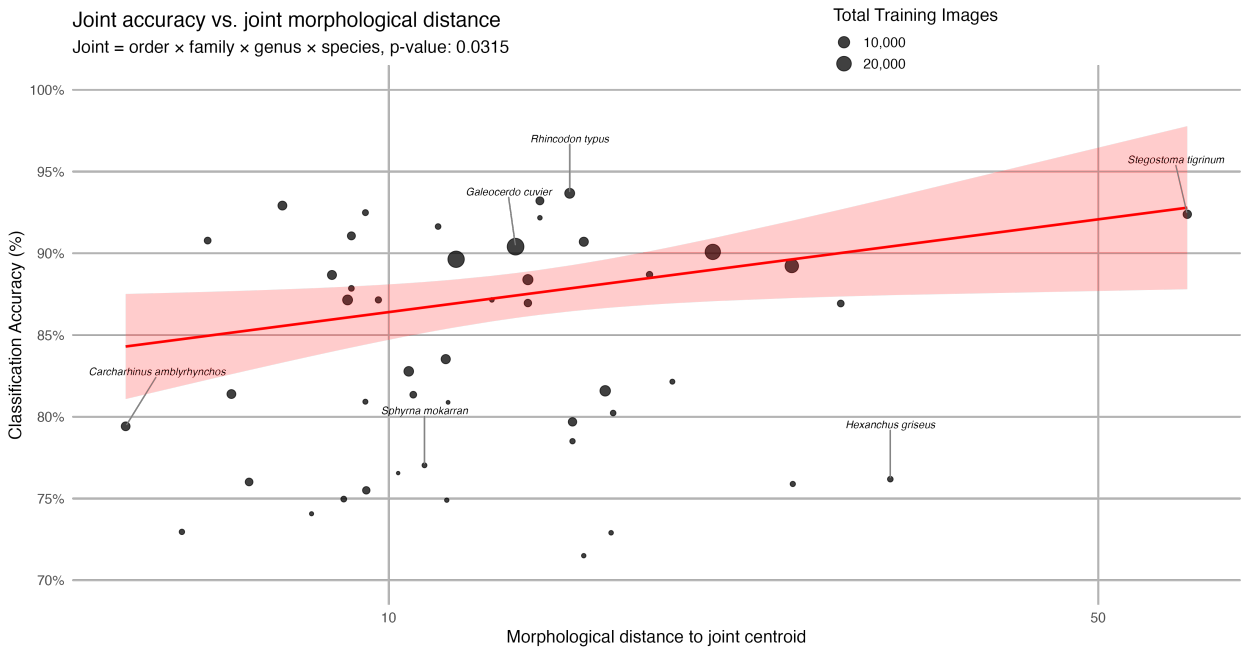


Figure D.1: Relationship between morphological distinctness and classification performance across sharks. Weighted linear regression shows that taxa with greater morphometric distance from their training centroids achieved higher $F_1$ scores.

## D.2 Taxonomic Summary

A critical step in developing the SD was establishing a minimum threshold for the amount of training data required to include a species in the classification pipeline. Within the sharkPulse training dataset, 228 shark species were represented by fewer than 200 labeled images, a level insufficient for stable model training. We therefore imposed a cutoff of 200 images per species, which reduced the immediate taxonomic coverage but greatly improved the reliability of predictions. After applying this threshold, 80 species remained with sufficient data representation to support robust classification. These species span 7 orders, 21 families, and 38 genera, forming the core taxonomic range of the current SD (version 5: Figure D.2). This cutoff not only ensures higher baseline accuracy but also highlights

priority taxa for future data augmentation, as continued ingestion of validated images will allow expansion of coverage beyond the present 80 species.

The phylogenetic distribution of classification performance provides an overview of how accuracy propagates across taxonomic levels. The circular tree illustrates $F_1$ scores for orders, families, and genera, represented by the performance of their branches, while species-level $F_1$ scores are indicated by colored terminal nodes (Figure D.3). Gray nodes mark species that currently lack sufficient training data to be classified directly at the species level, but that can still be reliably assigned to a higher rank. This hierarchical structure highlights the advantage of the SD approach: even when fine-scale resolution is unattainable due to limited data, the model can still provide ecologically meaningful identifications to 533 species total, nearly all shark species (excluding Pristiophoriformes, also known as sawsharks). Such flexibility is particularly valuable for ecological monitoring, where genus- or family-level identifications are often adequate for biodiversity surveys, community assessments, and conservation decision-making. By scaling its performance across taxonomic levels, the SD offers a practical and adaptive framework that balances data availability with classification accuracy.

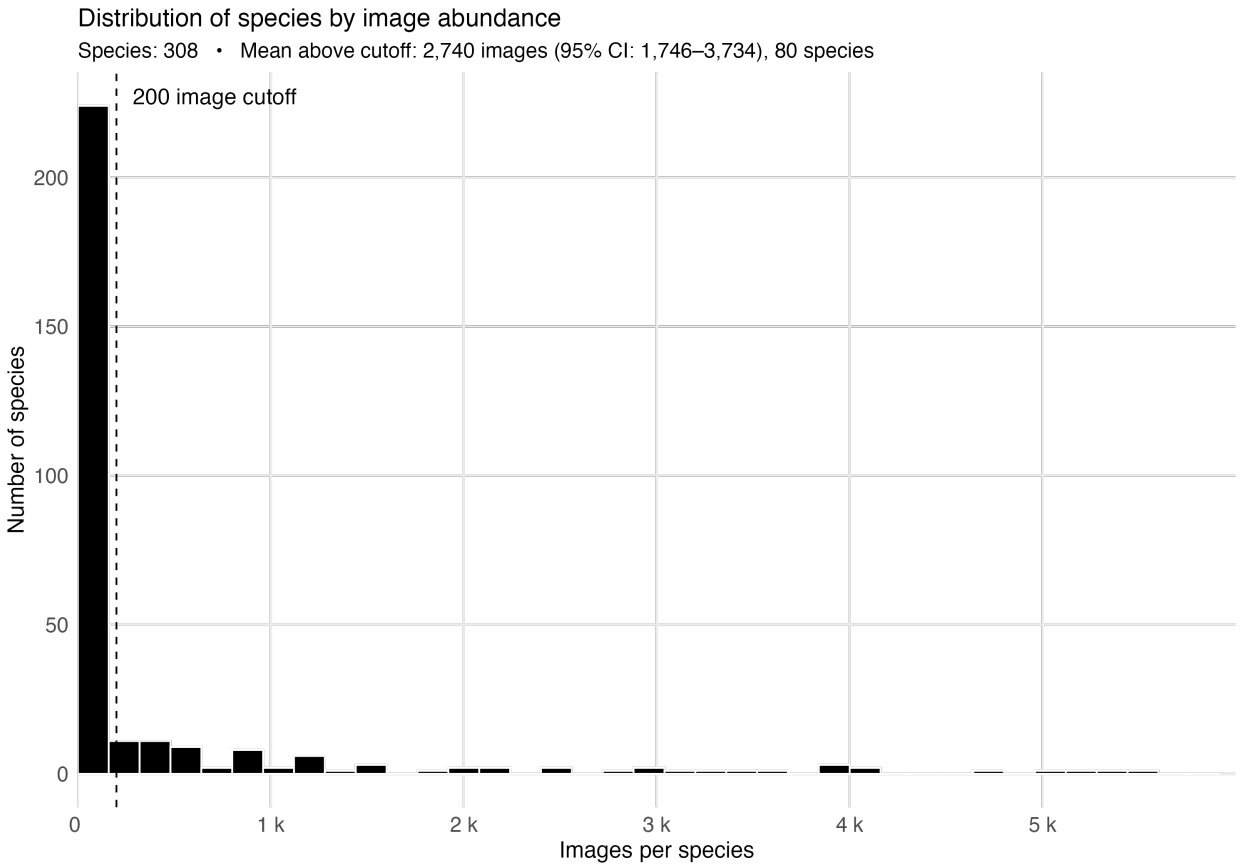


Figure D.2: Taxonomic coverage of the SDv5 after applying a 200-image training threshold. Eighty species spanning 7 orders, 21 families, and 38 genera are currently included, while species with fewer images remain candidates for future expansion.

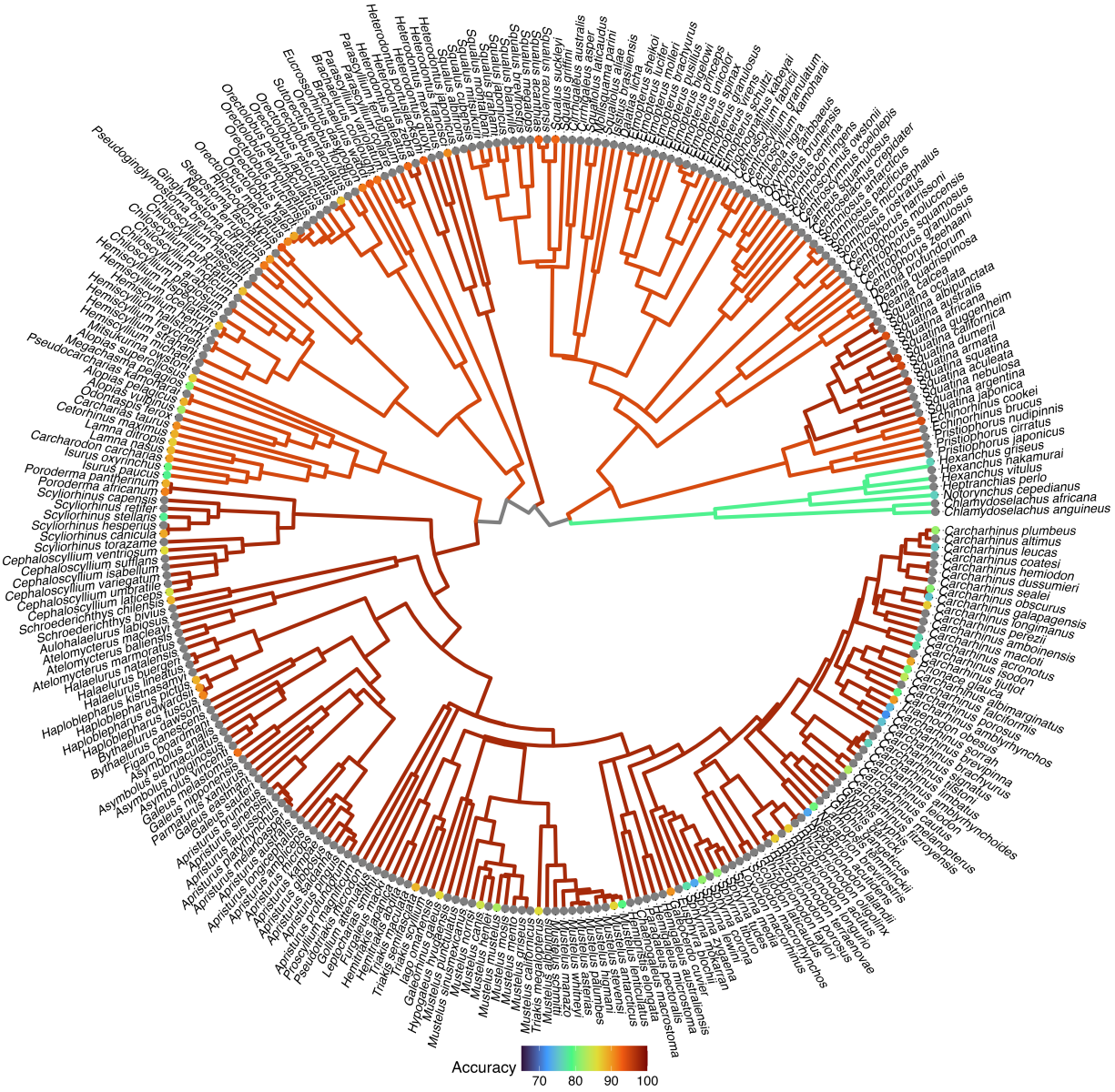


Figure D.3: Circular phylogenetic tree summarizing classification performance of the SDv5. Branch colors denote accuracies at order, family, and genus levels, while terminal nodes show species-level accuracy. Gray nodes represent species that can be classified only to higher taxonomic ranks.

# Appendix E

# BRUVs

This brief appendix summarizes contextual geographic and human population information, shark biodiversity, and infrastructural strategies to carry out the two Baited Remote Underwater Video (BRUV) surveys in the Hawaii and Palau regions.

The Hawaiian Archipelago lies in the Central Pacific and consists of 18 islands and atolls spanning 2,600 km, from Kure Atoll in the northwest to Hawai‘i Island in the southeast. It is one of the most remote and isolated archipelagos in the world. This volcanic chain includes the geologically younger and human-populated MHIs, as well as the older and largely uninhabited Northwestern Hawaiian Islands, which form part of a large marine refuge. The region hosts high levels of shark diversity, with reef-associated species (e.g., *Carcharhinus amblyrhynchos*) often abundant in atoll habitats, while pelagic species such as *Prionace glauca* and *Alopias pelagicus* migrate through offshore waters. Seasonal patterns in abundance and differences between reef and pelagic habitats make the archipelago a natural laboratory for studying shark ecology and conservation.

In the MHI survey, all sampling sites were separated by at least 500 m. Each BRUVs platform was equipped with an action camera (GoPro Hero models 5 and above), enclosed in either an acrylic tube (Blue Robotics, depth-rated to 100 m) for shallow deployments or an aluminum tube (Blue Robotics, depth-rated to 900 m) for deep deployments. The housings were mounted on galvanized iron frames and deployed from a motorized vessel, lowered to the seafloor by rope.

The Palau Archipelago is in the Western Pacific Ocean and consists of more than 340 islands spanning 200 km across southwest Micronesia. Its geological history reflects a dynamic interplay between volcanic activity and coral reef growth, producing a seascape of lagoons, barrier reefs, and marine lakes. Palau is considered a biodiversity hotspot, with reef shark populations (e.g., *Carcharhinus melanopterus* and *Triaenodon obesus*) particularly notable for their high densities compared to many other Pacific regions. In 2020, Palau strengthened its global conservation leadership by expanding the Palau National Marine Sanctuary to cover 470,000 km$^2$, representing 80% of the country's Exclusive Economic Zone (EEZ). This large-scale protection provides a critical setting to study how marine protected areas contribute to shark biodiversity, abundance, and habitat use.

## E.1 Video Processing Workflows

To complement the regional survey descriptions, we also evaluated the tradeoff between efficiency and accuracy across three video processing strategies: fully manual annotation, semi-automatic workflows incorporating partial automation with human review, and fully automated workflows using the SD pipeline with the SharkByte Graphical User Interface (GUI). Each approach was assessed in terms of annotation time and resulting classification performance based on a 6.9-hour subsample of footage from each survey, allowing us to quantify the balance between labor investment and predictive accuracy. Figure E.1 outlines these workflows schematically, with directional arrows indicating the progression from raw BRUV footage to annotated shark detections.

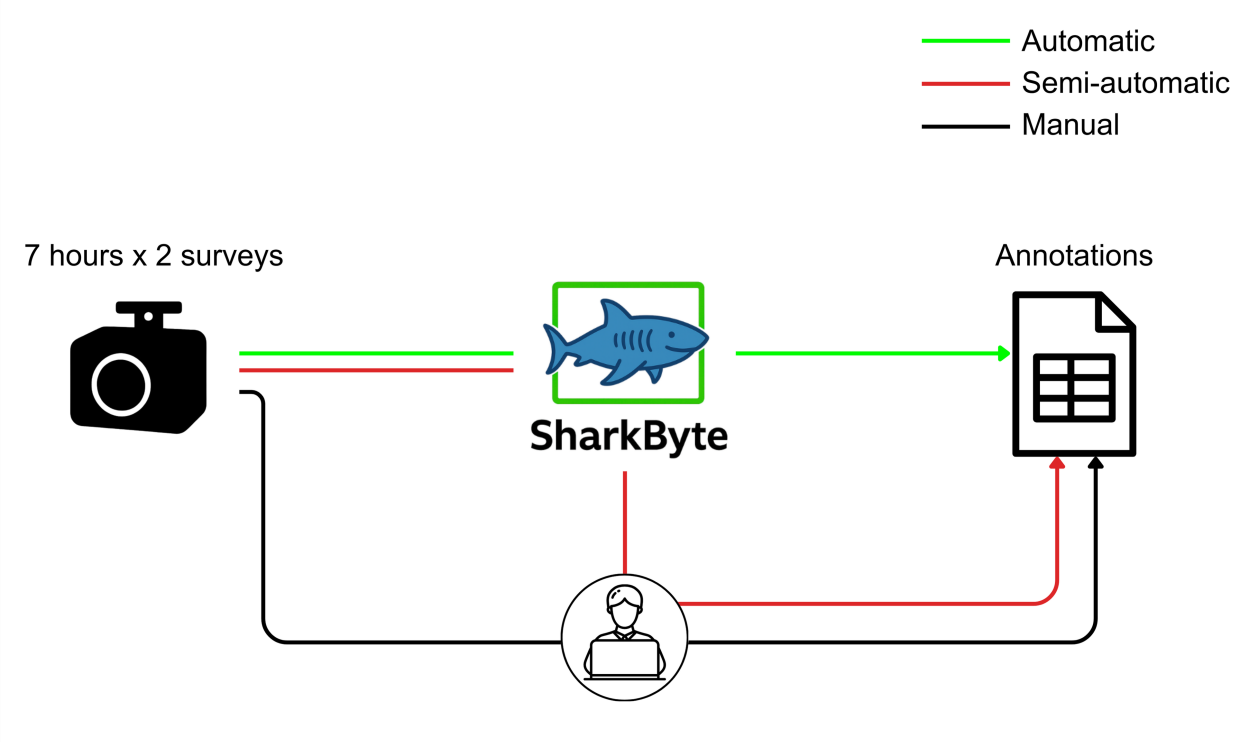


Figure E.1: Workflows for annotating BRUV footage. The black line represents manual annotation at real-time playback, the red line indicates a semi-automatic workflow combining automated detection with human review, and the green line shows the fully automated workflow using the SD pipeline with the SharkByte GUI. Directional arrows illustrate the progression from raw footage to annotated shark detections.

# Appendix F

# Instagram Sourcing Methods

Metadata were collected using the open-source tool InstaCrawlR [132]. We attempted to access the official IG Graph API through a Facebook Business Account, but business verification was not granted to academic users, preventing token generation. A third-party scraping platform (Apify) was then tested [155], which proved effective for structured tasks but required paid access and lacked transparency, limiting reproducibility.

To evaluate Apify's reliability, we implemented a workflow scheduled every two weeks to scrape the hashtag `#tigershark` for the preceding fortnight (December 21, 2023–May 1, 2024). This proactive schedule reduced risks of post expiration or deletion. While explicit evidence for post expiration remains unclear, short intervals facilitated ongoing monitoring and quality assurance milestones. The workflow was semi-automated but required regular maintenance.

The scheduled Apify-based workflow collected 3,843 posts for pilot validation, of which 498 (14.7%) were manually reviewed. From these, 398 posts (10.3%) describing eleven shark species were confirmed as unique, wild observations with geolocation and date metadata.

# Appendix G

# iNaturalist Diagnostic Plots

In this appendix, diagnostic plots were generated to evaluate the goodness of fit and predictive performance of the negative binomial models (Figures G.1, G.2, G.3, and G.4), using the tiger shark (*Galeocerdo cuvier*) in the Bahamas as an illustrative example, showing how well the fitted values capture observed variability and residual structure through time.

The diagnostic plots indicate that the continuous and point-estimate models adequately capture the general pattern of annual shark sightings, though some deviations remain in years with sparse data. Both models show residuals largely centered around zero, suggesting no major systematic bias, but with mild heteroskedasticity where fitted values are small, reflecting underdispersion in low-count years.

Observed versus predicted totals show that most years fall close to the one-to-one line, with the models slightly overpredicting recent years (e.g., 2024–2025), which likely reflects the limited number of observations in earlier time periods relative to later ones. Overall, the diagnostics support a reasonable model fit for trend inference, while highlighting expected uncertainty associated with years of low sampling effort or few observations.

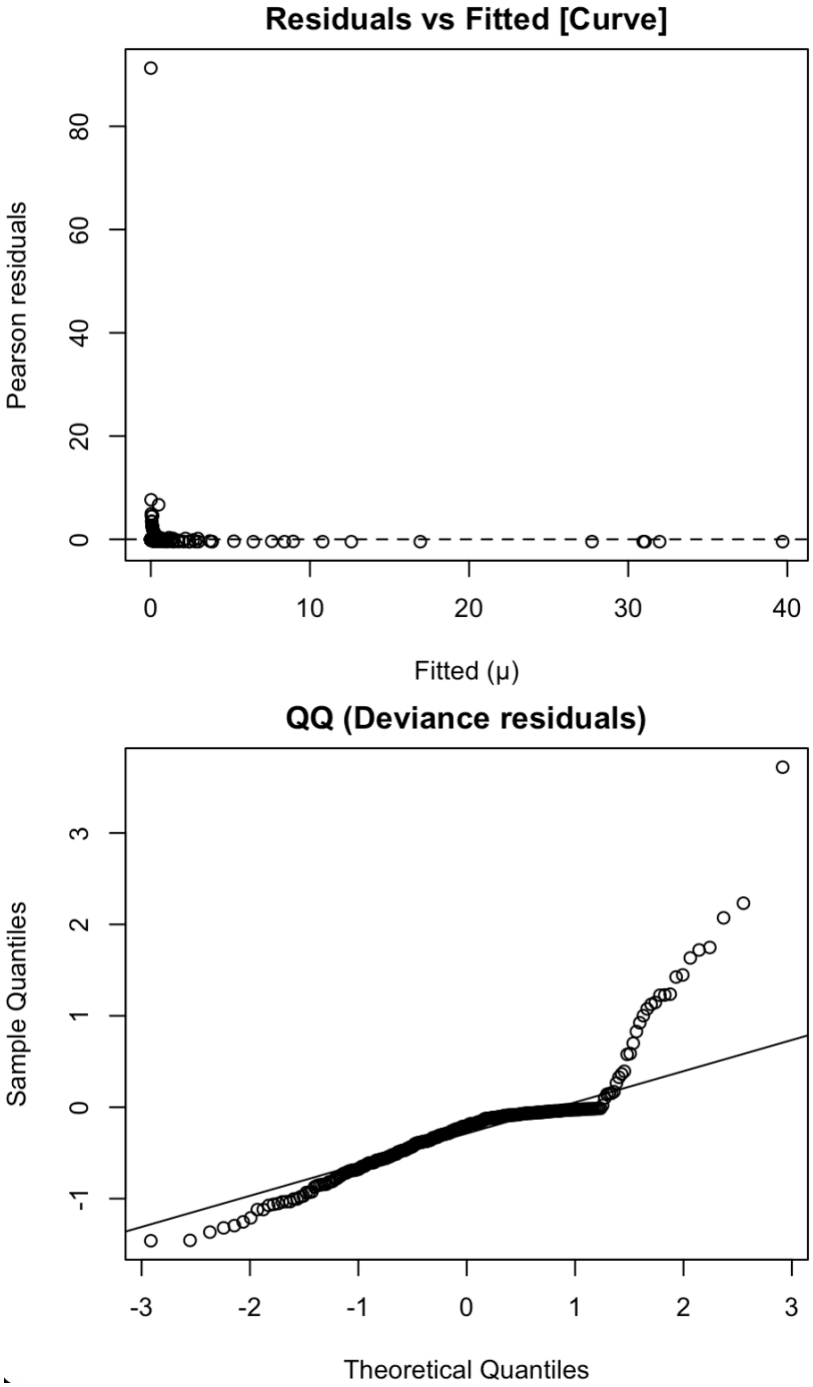


Figure G.1: Residual and goodness-of-fit diagnostics (top) for the continuous-year trend model, showing the relationship between fitted and residual values as well as the distribution of deviance residuals (bottom).

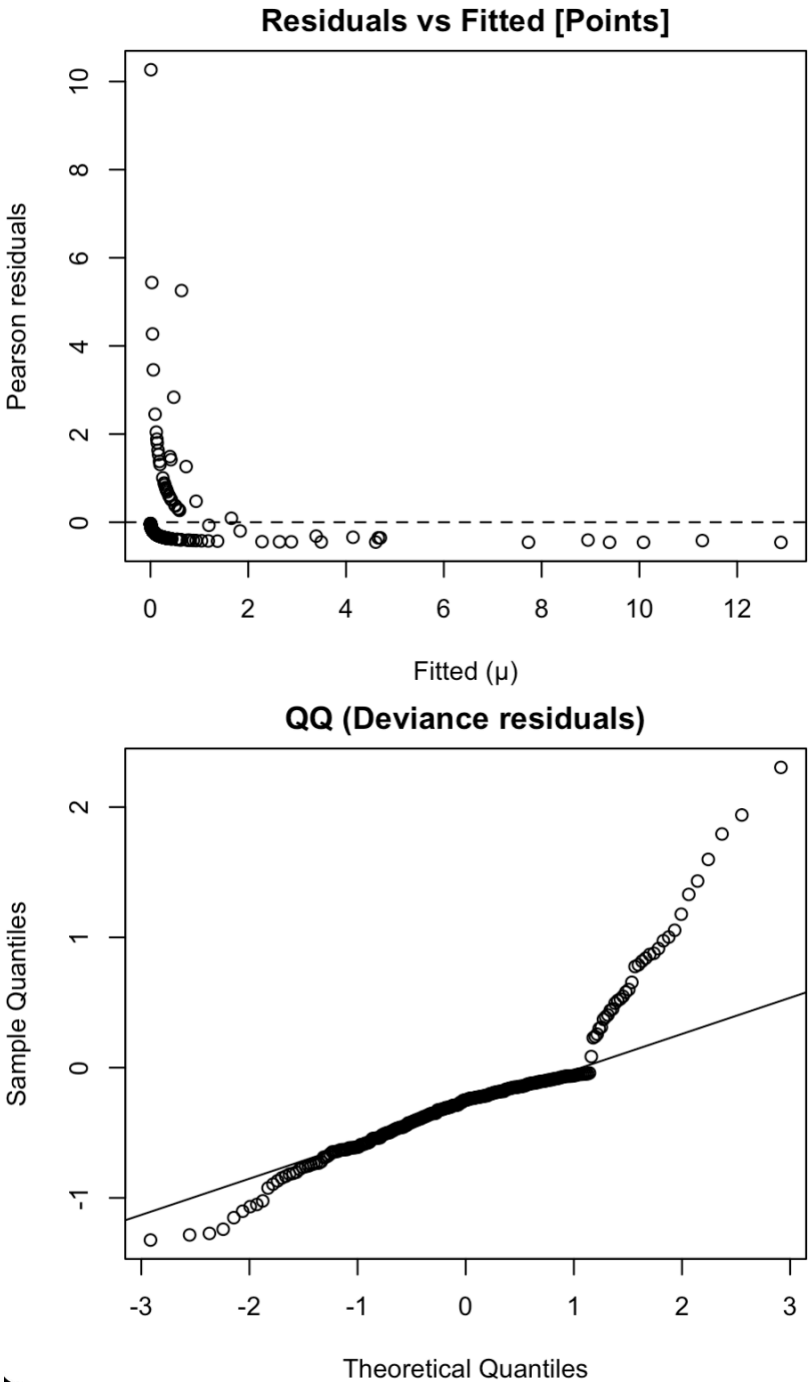


Figure G.2: Residual and goodness-of-fit diagnostics (top) for the point-estimate model, showing the relationship between fitted and residual values as well as the distribution of deviance residuals (bottom).

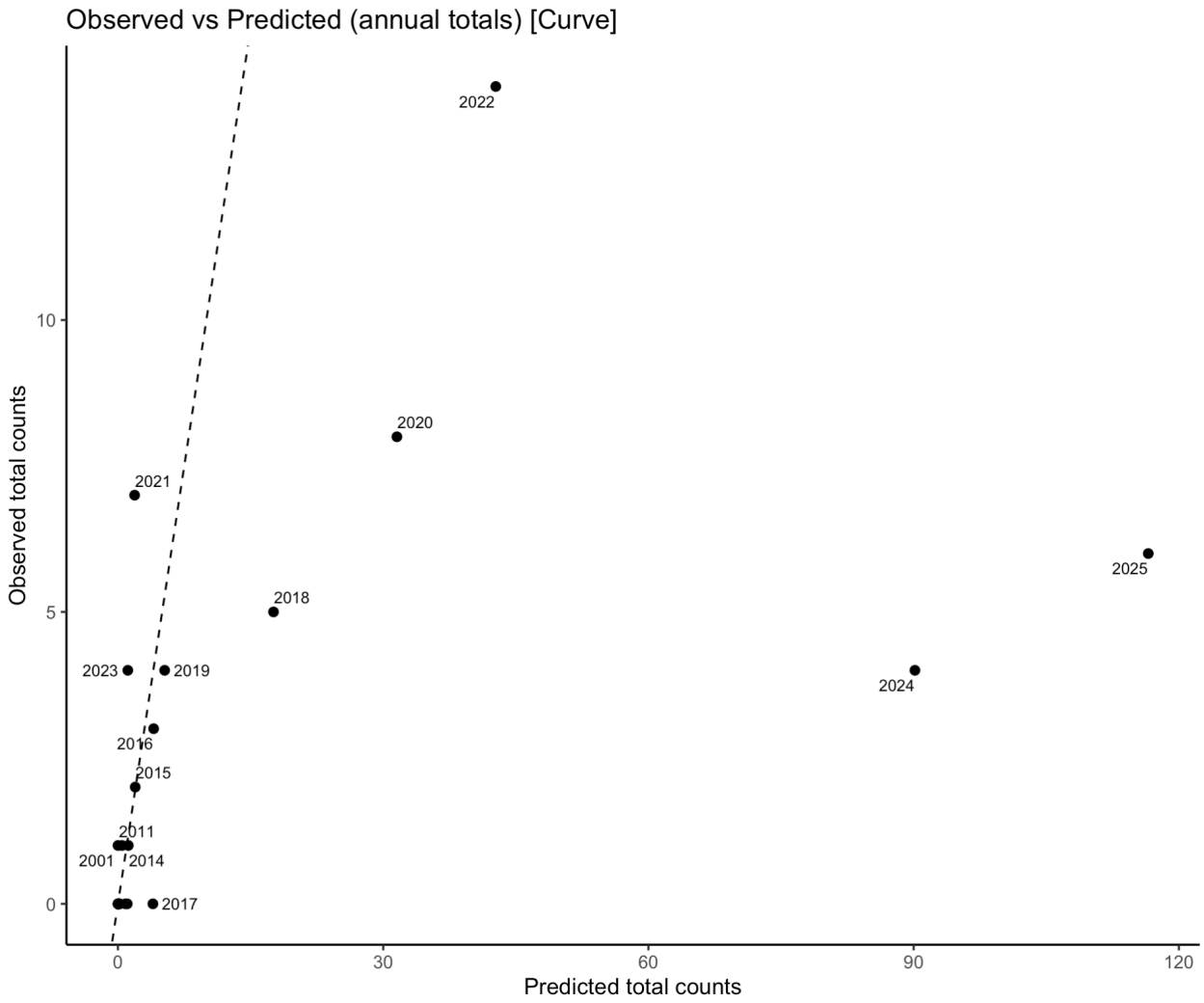


Figure G.3: Observed versus predicted shark counts at a monthly scale, illustrating the correspondence between model predictions and observed shark sightings.

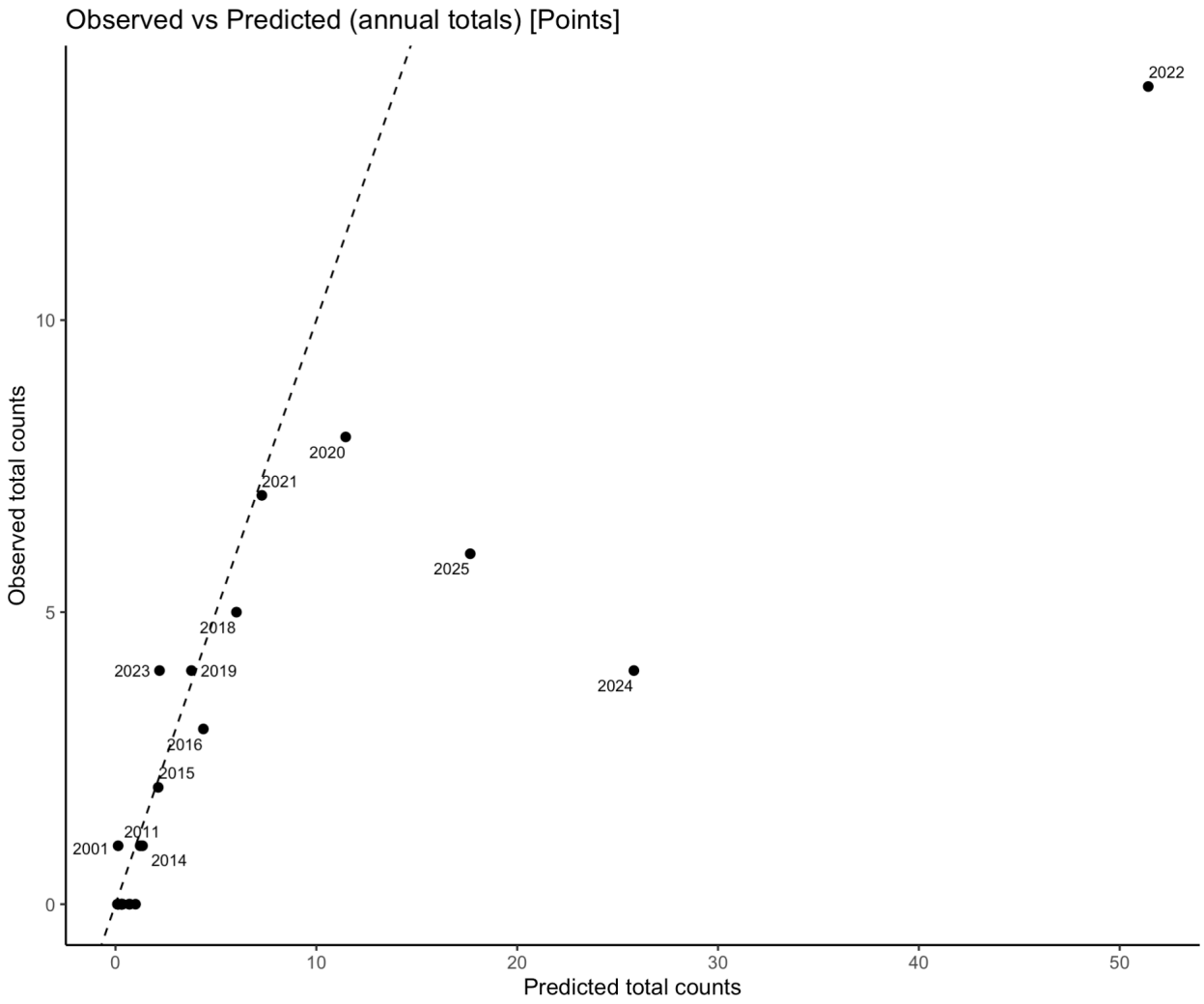


Figure G.4: Observed versus predicted shark counts at the annual scale, showing the correspondence between model predictions and observed shark sightings.