TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Afrasiyab Khalili 213856IASM

FROM PIXELS TO PATTERNS: AUTOMATED CLASSIFICATION OF FISH SWARMS IN UNDERWATER VIDEOS

Master's Thesis

Supervisor: Jeffrey Andrew Tuhtan
Tenured Associate Professor

Co-supervisor: Elizaveta Dubrovinskaya
Researcher

Tallinn 2024

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Afrasiyab Khalili 213856IASM

PIKSLITEST MUSTRITENI: KALAPARVEDE AUTOMATISEERITUD KLASSIFITSEERIMINE VEEALUSTES VIDEOTES

Magistritöö

Juhendaja: Jeffrey Andrew Tuhtan
Tenured Associate Professor

Kaasjuhendaja: Elizaveta Dubrovinskaya
Researcher

Tallinn 2024

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Afrasiyab Khalili

20.05.2024

Abstract

The migration patterns of fish are critical ecological phenomena, integral to managing and conserving aquatic resources. Accurately monitoring these patterns requires robust and precise classification systems. This study introduces an advanced computer-vision based approach to identify and classify fish migration behaviors focused on fish swarm activities from underwater video footage. Employing a novel data pipeline architecture known as 'Medallion,' this research streamlines the preprocessing and formatting of diverse video data sourced from the German Federal Institute of Hydrology. The approach involves deploying different preprocessing techniques tailored to specific behavior types: metadata matching for swarm class videos and a single fish model for isolating individual fish recordings.

Spatial feature extraction from detected fish bounding boxes, produced by a previously trained and validated Fish No-Fish model (YOLOv5), provides critical inputs for a Gradient Boosting Classifier tasked with behavior classification. Validation results show the model achieving an overall F1-score of 87%, with individual scores of 97% for fish swarm, 82% for single fish, and 80% for fish behavior, across a balanced dataset of 1851 videos. The test phase results further underscore the model's efficacy, especially in identifying the fish swarm behavior class with a precision of 97% and a recall of 93%.

The findings suggest that the integrated use of advanced machine learning techniques can significantly enhance the accuracy and efficiency of fish behavior analysis in ecological research and fisheries management. This study not only advances our understanding of fish behaviors in natural settings but also sets a new benchmark for technological applications in environmental conservation.

Keywords: Fish Behaviour Analysis, Video Classification, Machine Learning, Data Quality

The thesis is written in English and is 62 pages long, including 6 chapters, 28 figures, and 9 tables.

Annotatsioon

Pikslitest mustriteni: Kalaparvede automatiseeritud klassifitseerimine veealustes videotes

Kalade rändemustrid on kriitilised ökoloogilised nähtused, mis on veeressursside majandamise ja säilitamise lahutamatu osa. Nende mustrite täpne jälgimine nõuab tugevaid ja täpseid klassifitseerimissüsteeme. See uuring tutvustab täiustatud arvutuslikku lähenemisviisi kalade rändekäitumise tuvastamiseks ja klassifitseerimiseks veealuste videomaterjalide põhjal. Kasutades uudset andmejuhtmete arhitektuuri, mida tuntakse nimega "Medallion", lihtsustab see uurimus Saksamaa Föderaalsest Hüdroloogiainstituudist pärinevate erinevate videoandmete eeltöötlust ja vormindamist. See lähenemisviis hõlmab erinevate eeltöötlustehnikate kasutuselevõttu, mis on kohandatud konkreetsetele käitumistüüpidele: metaandmete sobitamine sülemiklassi videote jaoks ja üks kalamudel üksikute kalade salvestiste eraldamiseks.

Fish No-Fish mudeli (YOLOv5) abil loodud ruumiliste featuuride eraldamine tuvastatud kalade piirdekastidest annab kriitilisi sisendeid gradiendi võimendamise klassifikaatorile, mille ülesandeks on käitumise klassifitseerimine. Valideerimistulemused näitavad, et mudel saavutas 1851 videost koosnevas tasakaalustatud andmekogus üldiseks F1-skooriks 87%, kusjuures individuaalsed skoorid on kalaparve puhul 97%, üksikute kalade puhul 82% ja kalade käitumise puhul 80%. Katsefaasi tulemused rõhutavad veelgi mudeli tõhusust, eriti kalaparve mustrite tuvastamisel 97% täpsusega ja 93% tagasikutsumisega.

Tulemused viitavad sellele, et täiustatud masinõppetehnikate integreeritud kasutamine võib oluliselt suurendada kalade käitumise analüüsi täpsust ja tõhusust ökoloogilistes uuringutes ja kalanduse majandamises. See uuring mitte ainult ei edenda meie arusaamist kalade käitumisest looduslikes tingimustes, vaid seab ka uue võrdlusaluse keskkonnakaitse tehnoloogilistele rakendustele.

Märksõnad: kalade käitumise analüüs, video klassifikatsioon, masinõpe, andmete kvaliteet

Lõputöö on kirjutatud ingilise keeles keeles ning sisaldab teksti 62 leheküljel, 6 peatükki, 28 joonist, 9 tabelit.

List of Abbreviations and Terms

AB	Activity Bursts
AC	Average Confidence
AFC	Average Frame Coverage
BfG	German Federal Institute of Hydrology
CNN	Convolutional Neural Network
CLAHE	Contrast Limited Adaptive Histogram Equalization
CV	Centroid Variance
GBC	Gradient Boosting Classifiers
HPD	High-Performance Desktops
iFO	infrared Fish Observation
mAP	Mean Average Precision
MDS	Minimum Detection Size
MFC	Maximum Frame Coverage
PDD	Peak Detection Density
R-CNN	Region-Based Convolutional Neural Network
SFM	Single-Fish Model
SiamRPN	Siamese region proposal network
TDF	Total Detections per Frame
YOLO	You Only Look Once

Table of Contents

List of Figures

List of Tables

1. Introduction

Fish migration is a critical ecological phenomenon affecting environmental conservation and fisheries management [1, 2]. Accurately monitoring fish migration patterns and behaviours through video surveillance allows researchers and policymakers to make informed decisions that enhance conservation efforts and optimize fisheries management [3]. The traditional reliance on human manual observation for classifying and analyzing fish in video data poses significant challenges, including substantial labor costs and potential biases in data interpretation [4]. However, with developement of technology, the automated monitoring in ecosystem aims to decrease the reliance on manual observation, thus increasing the efficiency, accuracy, and scalability of data processing [5]. By employing advanced machine learning techniques to automate the detection and classification of fish behaviour, such as swarm activities, researchers can achieve a more consistent and objective analysis of underwater video data. This reduction in human labor not only cuts costs but also leverages computational precision to handle vast datasets that would be unmanageable manually [6].

Therefore, the development of an automated classification system for underwater fish videos not only supports significant ecological and economic outcomes but also represents a critical step forward in the application of artificial intelligence in environmental science. The combination of environmental management strategies with advanced technological approaches ensures the sustainability and effective stewardship of underwater resources.

Early efforts in underwater fish monitoring utilized basic video recording equipment and required manual review by experts, which was highly labor-intensive and prone to human error. With advancements in image processing and computer vision, automated systems began to gain popularity. Strachan [7] was among the first to apply computer vision techniques to identify marine species based on color and shape descriptors, paving the way for further technological advancements in underwater fish monitoring.

The development of automated fish counters introduced systems that used controlled illumination and stereo cameras to enhance the identification accuracy in constrained environments [8]. However, these systems struggled in natural, uncontrolled settings due to variable environmental conditions.

The progression in underwater video analysis for fish detection, particularly regarding the

integration of environmental adaptability, marks a significant evolution in the technologies employed. Fabic et. al [9] utilized shape analysis and blob counting techniques to identify and count fish. Their approach also involved some degrees of environmental normalization to mitigate issues such as shadows and reflections. However, these techniques required predefined settings and parameters that were not dynamically adjustable, limiting their effectiveness across different or changing underwater environments.

As first demonstrated by the ground-breaking work of Szegedy et al. [10], the introduction of deep learning has signaled a significant advancement in object detection, enabling the precise localization of objects across various classes. This technology has facilitated rapid advancements in underwater fish detection, employing deep learning techniques for various ecological monitoring and research purposes. Following these foundational developments, Zhao et al. [11] reviewed the effectiveness of deep learning in object detection, further establishing the robust capabilities of these models in complex scenarios, including underwater environments. Li et al. [12] leveraged a Fast Region-based Convolutional Neural Network (R-CNN) network to develop an automatic fish identification system. Their model, tested with the ImageCLEF dataset from the Fish4Knowledge project, achieved a mean Average Precision (mAP) of 81.4%, showcasing an 80 times faster detection rate compared to earlier R-CNN models. This progression was furthered by subsequent studies by the same team, which used this dataset to train on Faster R-CNN, achieving an mAP of 82.7%, and later developed a more efficient, lightweight neural network that improved detection capabilities [13, 14].

However, challenges persist, particularly concerning the quality and suitability of datasets for training these advanced models. Many publicly available datasets, such as Fish4Knowledge, do not capture the variability and complexity typical of underwater environments, often due to issues like low resolution and lack of environmental context [15, 16]. This limitation is evident in other popular datasets like those developed by Cutter et al. [17] and Anantharajah et al. [18], which suffer from similar issues of cropping and small sample sizes.

In response to the evolving needs of ecological monitoring, newer underwater monitoring systems have incorporated advanced technologies that improve the accuracy and efficiency of data collection in complex aquatic environments. The FishCam underwater observation system offers a versatile, open-source solution specifically designed for extended deployments in aquatic environments [19]. This system employs customizable configurations to adapt to various research needs, further enhancing the capabilities for detailed and prolonged ecological studies. These systems not only overcome some of the shortcomings of earlier technologies but also enhance the capability for long-term ecological monitoring,

providing crucial data that supports sustainable management and conservation efforts. Adding to these advancement, the infrared Fish Observation (iFO) system [20], utilizes low-cost, open-source infrared video systems to monitor aquatic life behaviour without the disruptive impact of traditional lighting, essential for accurate nighttime observations. Similarly, the Riverwatcher system (used in this work) [21], leverages infrared scanning technology to passively and non-invasively monitor fish migrations in rivers and fishways. These systems not only overcome some of the shortcomings of earlier technologies but also enhance the capability for long-term ecological monitoring, providing crucial data that supports sustainable management and conservation efforts.

These advancements represent a significant step forward in the deployment of environmental near real-time monitoring technologies, shifting from reliance on imperfect datasets to real-time quality dataset, adaptable solutions that can handle the inherent challenges of aquatic environments.

Recognizing dataset limitations and improved near real-time aquatic observation systems, recent efforts by Cai et al. [22] and Wang et al. [23] have focused on developing tailored datasets and employing state-of-the-art models that better reflect the real-world conditions of underwater habitats. These models have shown promising results, though the challenge of applying them across different underwater settings remains, due to variations in water clarity, illumination, and environmental complexity.

While advances in computer vision and machine learning have significantly enhanced the capabilities of automated fish counters, most applications continue to rely on High-Performance Desktops (HPD) due to their computational power. This dependence presents a significant barrier for real-time, field-deployable systems, as the complexity of these methods often precludes their use on low-cost, low-power embedded hardware, which are crucial for widespread and ubiquitous monitoring in diverse environmental conditions. The challenges associated with implementing such technology in the field, due to limited computational resources and the technical difficulties of deployment, highlight a significant gap in the current research landscape [24, 25].

Addressing these limitations, the integration of high-performance computing and embedded systems into fish detection technologies has been crucial. Jürgen Soom's recent work [26] represents a significant leap in addressing the variability of natural underwater environment by introducing an environmentally adaptive multi-stage classification process where the methodology allows for real-time, robust classification under different conditions such as turbidity and variable lighting [27, 28]. These studies demonstrate the evolution from high-cost, high-maintenance setups to more sustainable, low-power systems that

can perform complex analyses. This innovative approach not only facilitates the robust classification of videos with and without fish but also categorizes commonly occurring environmental conditions that impact visibility and detection accuracy, such as turbidity and light overexposure. The dual capability allows for enhanced adaptability and reliability of fish monitoring systems in variable freshwater environments. Comparatively, frame differencing emerged as the more effective technique in his study, achieving a mean accuracy of 88.4%, versus the 82.1% accuracy obtained through scanlines. These methods offer substantial improvements over traditional systems [24], even more optimized systems, such as an embedded fish counter on a Raspberry Pi with controlled environments, have reached accuracies up to 98%. This contrast illustrates the complexities of applying these technologies in wild, unstructured environments where the performance often degrades, with F1-scores potentially dropping below 50% [29].

In exploring the realm of fish behaviour analysis, according to the research carried out by Hu Jun et al. [30], innovative methods are developed to monitor fish behaviour effectively within aquaculture environments. Utilizing a low-cost underwater imaging system paired with an enhanced version of the You Only Look Once (YOLO) V3-Lite deep learning model, the research captures and analyzes various fish behaviours based on visual cues in the images, such as movement patterns during hypoxia, changes in posture or activity during feeding, and normal swimming behaviours and enables rapid processing, allowing behaviours to be detected and analyzed in real time, which is vital for timely adjustments in aquaculture management. Preprocessing techniques like contrast enhancement and noise reduction refine the visual data, facilitating precise behaviour classification through real-time image processing. He Want et al. [31] further advanced fish behaviour monitoring by identifying abnormal behaviours in fish within recirculating aquaculture systems, tackling the difficulties posed by small target sizes and occlusions due to high-density fish populations. By refining the YOLOV5 model for greater accuracy and combining it with the robust Siamese region proposal network ++(SiamRPN++) for tracking, their integrated system offers substantial improvements in both speed and detection accuracy. These advancements not only enhance real-time monitoring capabilities but also bolster proactive health management strategies, potentially leading to significant economic benefits by minimizing the risks of fish mortality in aquaculture practices. Further extending these capabilities, in the study of Iqbal et al. [32] research provides a significant tool for optimizing feeding strategies, thereby reducing waste and contributing to more sustainable fish farming practices by enabling precise and timely feeding decisions. A custom convolutional neural network (CNN) classifies fish behaviour into normal and starvation states with a demonstrated accuracy of 98%. This model, tested on a dataset of 2000 images from video footage of black scraper fish, incorporates three fully connected layers and max-pooling to enhance its predictive capabilities, highlighting the model's

success in using visual data to discern behavioural patterns that indicate starvation.

Building on the motivation rooted in technological advancements for ecological monitoring, the role of institutions like the German Federal Institute of Hydrology (Bundesanstalt für Gewässerkunde – BfG) becomes indispensable. BfG not only spearheads research on hydrological, ecological, and geoscientific issues but also emphasizes the sustainable management of water resources and the protection of freshwater ecosystems. Their extensive research activities provide crucial data that supports a wide array of ecological assessments and advancements.

In this research endeavor, BfG's initiatives align closely with our objectives, particularly their innovative use of technology in environmental monitoring. As we delve deeper into the specifics of fish behaviour analysis through video classification, the foundation laid by BfG's diverse hydrological research will be crucial. This backdrop is not just about leveraging a dataset but understanding how such data are gathered and utilized to drive significant environmental outcomes. By situating our study within the framework of BfG's broader research activities, we underscore the synergy between academic research and practical, real-world application in ecological management.

Recognizing previously referred challenges and lack of research study on fish behavioural analysis, this research aims to harness the power of cutting-edge machine learning technologies to revolutionize the way we understand and interact with marine life. The objectives of this study are designed to address specific gaps in current monitoring techniques by introducing an automated, robust, and scalable system for classifying fish swarm behaviours from underwater video data. This system promises not only to improve the precision of ecological assessments but also to enhance the efficacy of conservation efforts and fisheries management. With these goals in mind, the following objectives have been set to guide the research:

Primary Objective:

- To develop and validate an automated classification system capable of accurately identifying fish swarm behaviours from underwater video footage.

Secondary Objectives:

- To implement and assess the efficacy of advanced machine learning techniques, specifically Gradient Boosting Classifiers, in distinguishing between swarm and non-swarm activities in underwater environments.
- To leverage a robust feature extraction methodology that utilizes spatial features

derived from video frame analysis, ensuring detailed and accurate input data for model training.

- To conduct rigorous model evaluation through methods such as k-fold cross-validation and train-test splits to optimize and verify the model's performance across various metrics including accuracy, precision, recall, and F1 scores.
- To refine the model's predictive capabilities through systematic hyperparameter tuning using GridSearchCV, focusing on parameters like the number of estimators, learning rate, and tree depth to achieve the best classification results.

This study consists of five chapters. Chapter 2 presents the background, detailing the fish migration behaviours. Chapter 3 provides the brief information about the data sources, methodologies, and theoretical frameworks employed, including Medallion architecture for data preprocessing. Chapter 4 discusses the results, showcasing the dataset preparation results, the performance of the Fish-No Fish (FNF) model and the Gradient Boosting Classifier in classifying fish swarm behaviours. Chapter 5 provides a thorough conclusion to the investigation, summarizing the findings, and outlining potential future research directions.

2. Background

2.1 Migration Behaviour Classes

In the forthcoming subsections, we delve into the nuanced definitions that categorize each class of fish migration behaviour captured in video, the process of data preparation, and the array of challenges encountered while curating the final dataset. The study of fish migration classes through video analysis has delineated three primary behaviours (Fig 1) that offer insights into the piscine lifecycle within their underwater domain, which are:

- **Fish Migration Behaviour** - this category is used for videos featuring single fish, each showcasing one of six distinct migration behaviours through the underwater counter, defined as the sub-classes UP-IN, DOWN-IN, UP, DOWN, IN-DOWN, IN-UP.
- **Fish** - a category reserved for videos depicting single fish instances in which a fish appears in the video, but fails to exhibit any of the six migration behaviours. This class is representative for fish which remain in the counter for the duration of the video and do not migrate into or out of the camera (dweller). In addition, this class is also used for fish which partially enter the top, bottom, left or right side of the camera but do not swim into it fully, thus not migrating into, through or out of the camera (drop-out).
- **Fish Swarm** - this grouping comprises footage of multiple fish seen simultaneously in the video, with a threshold set to five fish per frame as the working definition of a fish swarm.

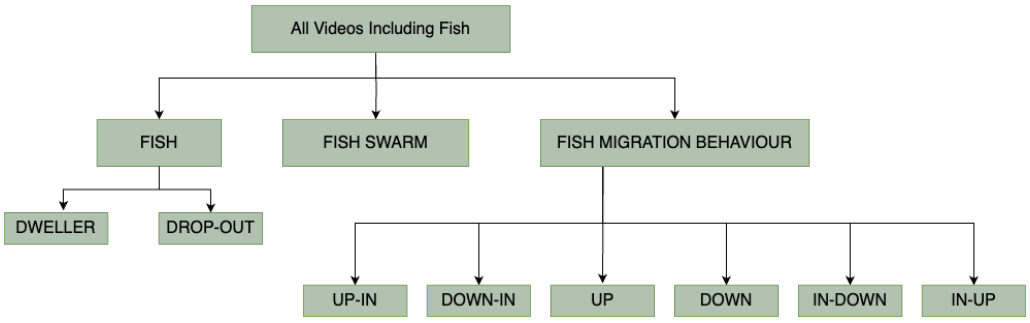
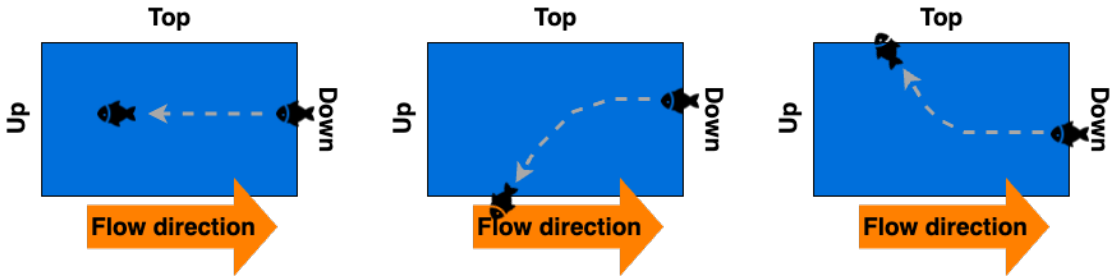


Figure 1. Hierarchical Classification of Fish Behaviours.

2.1.1 Fish Migration Behaviour

Fish migration behaviour videos are sub-classified into six different types based on the fish behaviour during the duration of a single video:

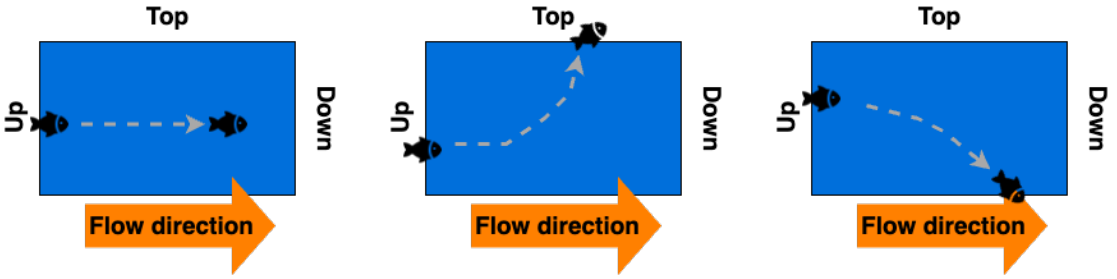
1. **UP-IN** - this behaviour has three sub-fish behaviours which provide the full definition:
   - Fish enters the view from the downstream end and does not migrate out of view (Fig 2a).
   - Fish swims upstream (enters the view from the downstream end) but does not migrate out of view or swims downwards to the bottom camera field (Fig 2b).
   - Fish swims upstream (enters the view from the downstream end), does not migrate out of view or swims upwards to the top camera field (Fig 2c).



(a) Fish entering the view (b) Fish swims upstream (c) Fish swims to the top field

Figure 2. Different types of UP-IN behaviour.

2. **DOWN-IN** - This behaviour has three sub-fish behaviours, which provide the full definition:
   - Fish enters the view from the upstream end, does not migrate out of view (Fig 3a).
   - Fish swims downstream (enters the view from the upstream end), does not migrate out of view or swims upwards to the upper camera field (Fig 3b).
   - Fish swims downstream (enters the view from the upstream end), does not migrate out of view or swims downwards to the bottom camera field (Fig 3c).
3. **UP** - this behaviour occurs when fish leaves the image and has the following definition:
   - Fish enters the view from the downstream end, migrates out of view upstream (Fig 4).



(a) Fish entering from upstream (b) Fish swims downstream (c) Fish swims to bottom field

Figure 3. Different types of DOWN-IN behaviour.

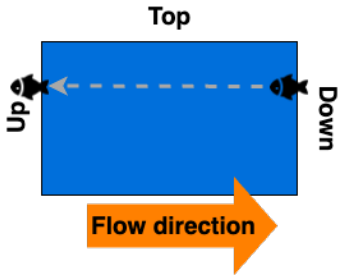


Figure 4. UP behaviour - Fish entering from downstream end.

4. **DOWN** - This behaviour occurs when the fish leaves the image and has the following definition:
   - Fish enters from the upstream edge and leaves through the downstream edge (Fig 5).

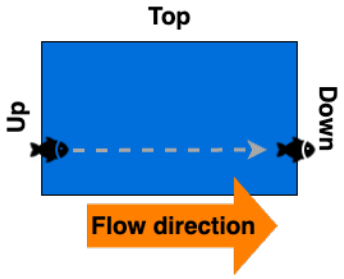
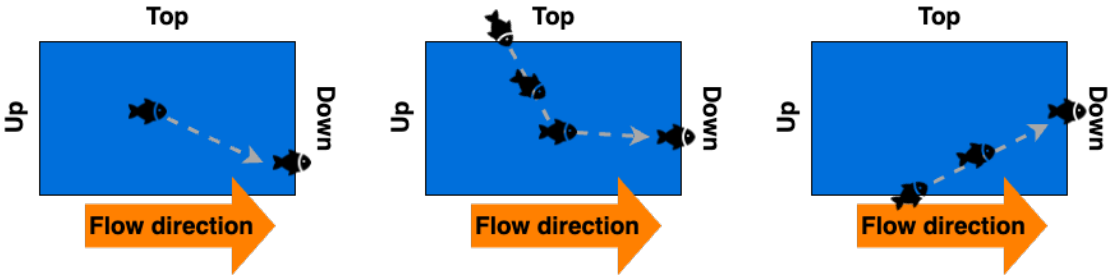


Figure 5. DOWN behaviour - Fish entering from the upstream end.

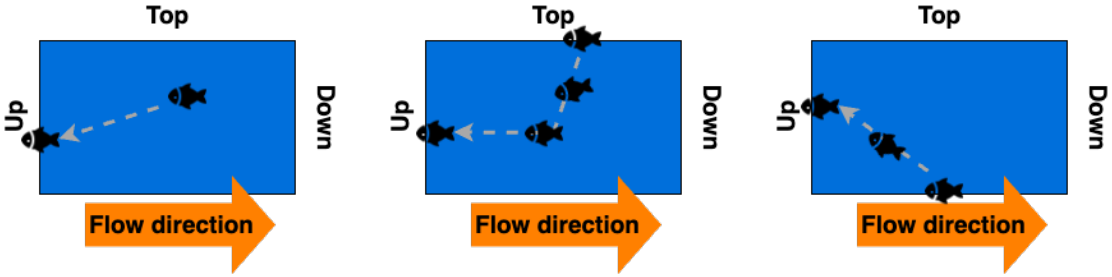
5. **IN-DOWN** - this behaviour occurs when the fish leaves the image and has the following definition:
   - Fish is in view at the beginning of the video and migrates out of view downstream (Fig 6a).
   - Fish enters from the top and migrates out of view downstream (Fig 6b).
   - Fish enters from the bottom and migrates out of view downstream (Fig 6c).
6. **IN-UP** - this behaviour occurs when the fish leaves the image and has the following definition:



(a) Fish in the view at start (b) Fish enters from top (c) Fish enters from bottom

Figure 6. Different types of IN-DOWN behaviour.

- Fish is in view at the beginning of the video and migrates out of view upstream (Fig 7a).
- Fish enters from the top and migrates out of view upstream (Fig 7b).
- Fish enters from the bottom and migrates out of view upstream (Fig 7c).



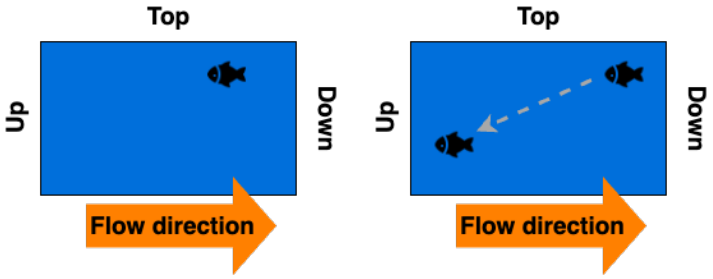
(a) Fish in the view at start (b) Fish enters from top (c) Fish enters from bottom

Figure 7. Different types of IN-UP behaviour.

2.1.2 Fish

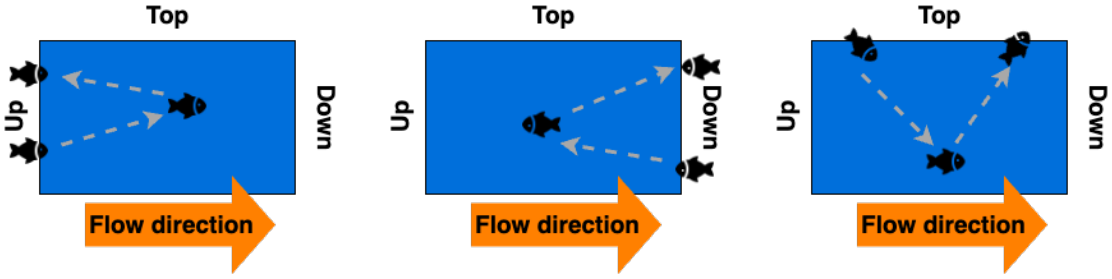
Fish videos are mainly categorized into two different types based on the fish behaviour during the lifecycle of the video, the behaviour types include:

1. **DWELLER** - this behaviour occurs when the fish is in view at the beginning (Fig. 8a) and at the end of the video (Fig. 8b). It shows all kinds of movements or position changes won't be a limitation for this type of behaviour.
2. **DROP-OUT** - This behaviour occurs when the fish is entered the view from upstream (Fig. 9a), downstream (Fig. 9b) or top (Fig. 9c) and leaves the video from upstream, downstream or top, respectively.



(a) Fish in the view at start (b) Fish in the view at end

Figure 8. Example of DWELLER behaviour.



(a) Fish entering the view from upstream (b) Fish entering the view from downstream (c) Fish entering the view from top

Figure 9. Different types of DROP-OUT behaviour.

2.1.3 Fish Swarm

Fish Swarm videos are classified when five or more fish pass through a video frame in groups (Fig. 10).

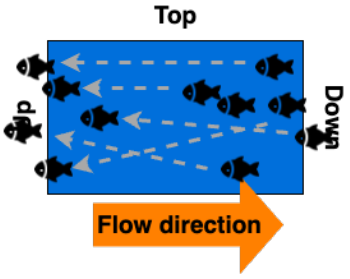


Figure 10. Fish swimming direction shown for emphasis. Swarm videos can include fish swimming upstream or downstream and frequently include fish swimming in both directions.

3. Methods

3.1 Data Pipeline: Medallion Architecture

The dataset at the core of this research represents a comprehensive collection of underwater video footage, specifically tailored to study and analyze fish migration patterns. This is a private dataset, and the right to use it was kindly given by The German Federal Institute of Hydrology (BfG) as this study is part of a collaboration project between a research group at TalTech and the BfG. The dataset for this study was prepared to facilitate the binary classification of underwater videos into three primary categories: FISH SWARM and NON FISH SWARM (FISH OR FISH BEHAVIOUR). The research provide an account of the methodological approach undertaken to curate a diverse dataset, aiming for an equitable distribution across individual fish, fish swarm, and varied fish behavioural videos. These have been methodically gathered from three different river locations in Germany to feed into the data classification framework. The dataset amalgamation involved selecting video footage from trio of sites, each chosen to reflect the heterogeneity of aquatic settings, thereby ensuring an all-encompassing dataset that captures an array of complex environmental conditions under which fish operate.

The Medallion architecture is a multifaceted framework utilized to transform and transition raw video data into an analytically viable and structured format. At its core, it integrates data warehousing methodologies with the robustness required for big data processing, particularly in the domain of ecological video data analysis [33, 34]. This architecture is integral for datasets that demand a sequence of refined transformations, each building upon the last to gradually enhance data quality and relevance to the research questions at hand [35]. The architecture's first layer is the acquisition of raw video data (see Fig[11), which is mainly split into three sublayers, which are:

1. 2019 Fish Swarm data pipeline (highlighted with a red dashed line).
2. 2020 Fish Swarm data pipeline (highlighted with a red dashed line).
3. 2020/2021 Fish & Fish Behaviour data pipeline (highlighted with a blue dashed line).

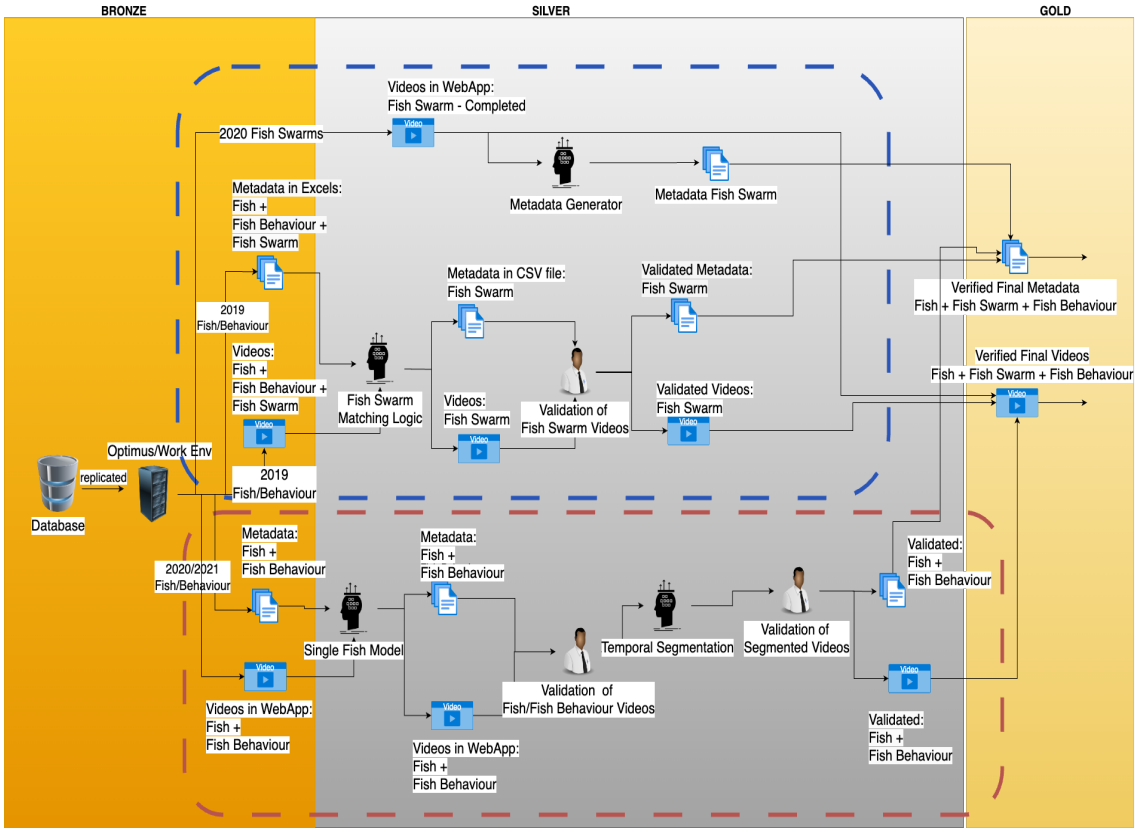


Figure 11. Data Processing Flow in Medallion Architecture.

3.1.1 Fish Swarm

For the Koblenz site's 2019 dataset, the acquisition is a gathering of extensive and unprocessed video files, collected over varied times and under diverse environmental conditions, representing a snapshot of underwater life in motion. These videos, each a temporal window into underwater ecosystems, are stored in their native high-fidelity format, capturing details that range from subtle fish behaviours to complex swarm dynamics.

The data, once collected, is replicated from its primary location, the Isis server, to the working environment, Optimus. This initial replication is a careful mirroring process that ensures data integrity and fidelity, providing a solid foundation for the forthcoming analytical operations.

Upon replication, we encounter the necessity for preprocessing - a bridge between raw data and analytical readiness. For the 2019 dataset, this involves two primary elements:

- **Metadata** - for the videos is initially scattered across several Excel files, each corresponding to a particular month of data collection. The files are named system-

atically (e.g., March2019_lower.xlsx, April2019_lower.xlsx) in Fig[12], with each containing vital descriptors such as date, time, species, size, and other observational details.

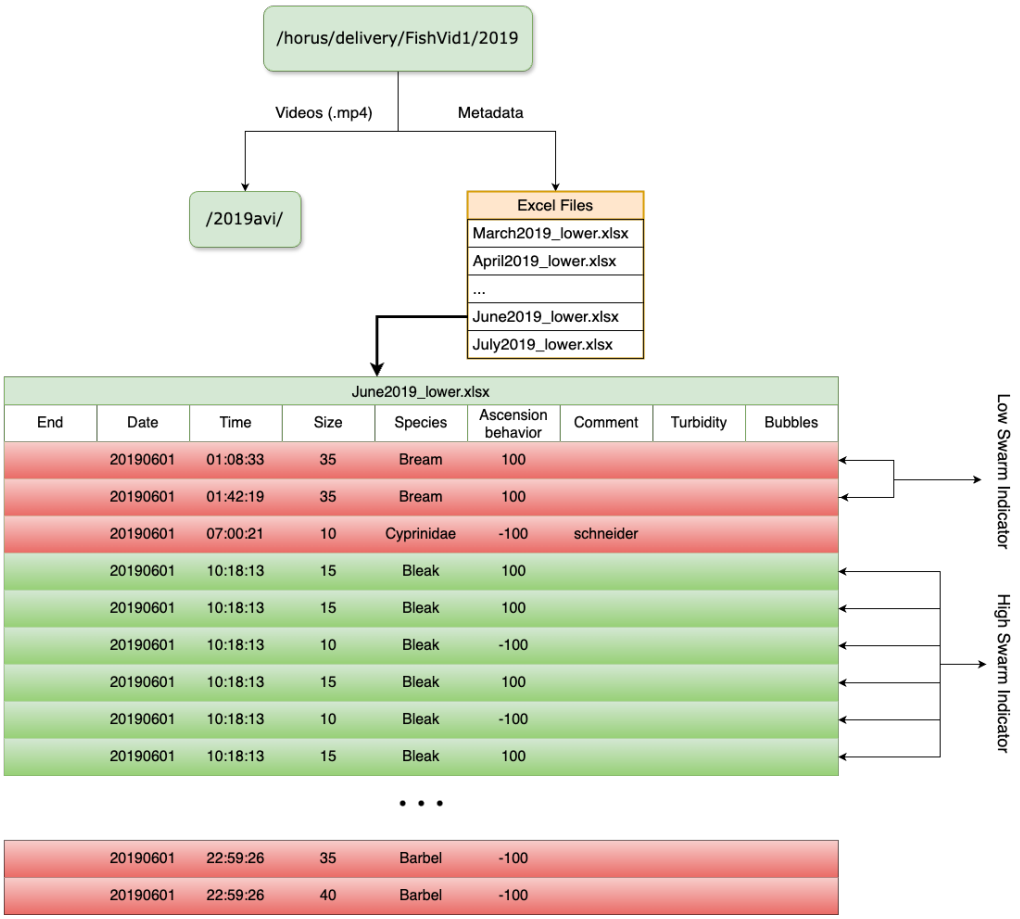


Figure 12. A sample of Fish Swarm 2019 June Raw Metadata Format.

The metadata's granularity is crucial for subsequent matching processes with the video files.

- **Video Files** - The video files are organized within the 2019avi directory (Fig[13]), sorted into daily subfolders denoted by date labels. The diversity within these files is vast, documenting various species, behaviours, and swarm events.

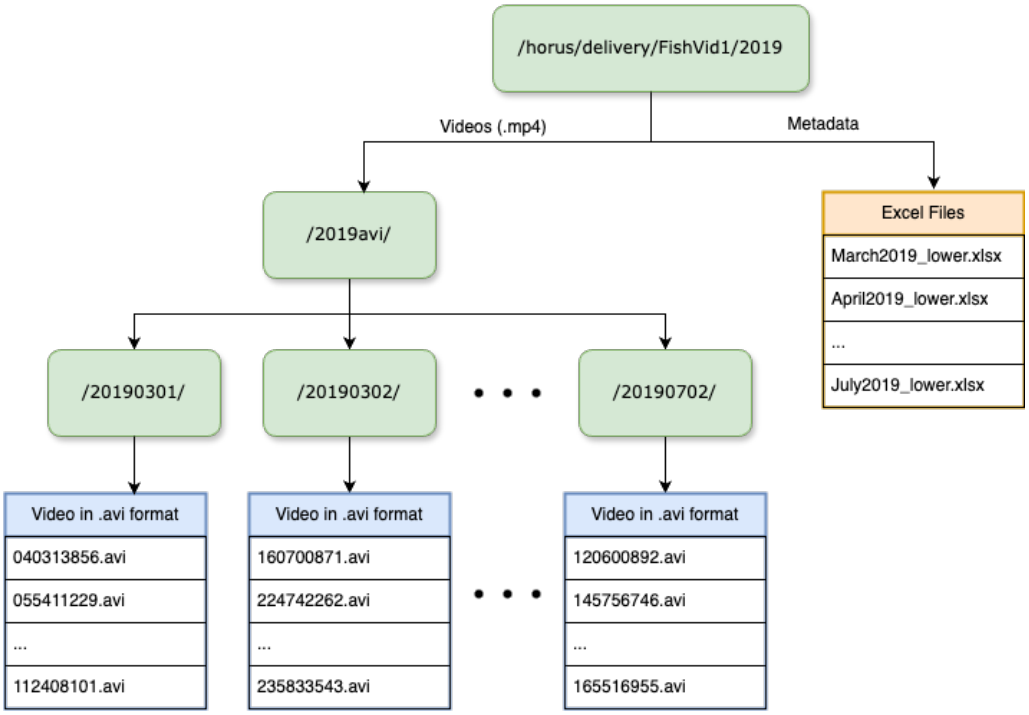


Figure 13. A sample of 2019 Koblenz dataset folder structure.

Following initial data structuring, the next significant endeavor is the consolidation of metadata. This step is crucial in transitioning from a dispersed to a centralized metadata framework. Here, individual Excel files are programmatically ingested into a singular data frame, with an emphasis on maintaining data coherence by selectively importing relevant columns.

Data Transformation and Preparation - the initial step within this phase involves the standardization of temporal data, necessitated by the disparate formats of date and time across the observational records. The standardization process harmonizes these records by merging the separate columns for date and time ('Date' and 'Time') into a singular datetime column ('Timestamp'). To enhance the matching process with the video files, the timestamps are further processed through a rounding operation. This operation modifies the precise seconds recorded in the timestamps down to the nearest minute, thus synchronizing with the video filenames that denote recording times by minute intervals (Table [1]).

The transformation from specific date and time columns to a unified datetime representation enables the system to treat time as a continuum, rather than as disjointed units. This continuity is crucial for any temporal analyses that follow, where events are contextualized within the flow of time rather than as isolated points.

Filtering and Isolation of Significant Events - Subsequent to the temporal standardization

Table 1. Transformation of Dataset Columns.

Original Columns	**Original Values**	**Transformed Values**
Date	25-03-19	25-03-19
Time	01:04:15	01:04:15
Species	Roach	Roach
Count	7	7
Timestamp	deafult_unknown	2019-03-25 01:04:15
Timestamp_rounded	deafult_unknown	2019-03-25 01:04:00

is the application of a set of filters—defined by ecological behavioural standards—to the data. These filters aim to isolate significant swarm events based on specific criteria that have been established through prior ecological research (Table[2]). In the context of fish swarm analysis, the main indicators for significant events include:

1. A threshold for the minimum number of fish records—events with counts equal to or greater than five fish are considered potential swarm activities.
2. A temporal proximity constraint—only those events for which the time span (the difference between the earliest and latest timestamps) does not exceed a three-minute threshold are retained. This constraint is predicated on the established behavioural patterns of fish swarms that tend to display significant activity within such a temporal window.

Table 2. Dataset Reduction after Fish Swarm Indicator Application.

Filter Criteria	**Entries Before**	**Entries After**
Count >= 5 and Max Timestamp - Min Timestamp < 3 mins	1865	306

Green highlighted records in Fig[12] are the example of possible fish swarm occurance since number of records are bigger than five fishes, time interval is between given threshold interval and should be classified as 'FISH_SWARM'.

This phase results in a dataset pruned and shaped by the precise requirements of the research study. The standardization of time and the strategic filtering of events serve as preludes to the subsequent phase of video matching. Through these processes, the data is transformed from its raw state into a format ready for the rigorous matching algorithm that follows, setting the stage for accurate and efficient swarm identification.

Swarm Identification and Video Matching - this stage introduces a rigorous matching algorithm that utilizes the final predicted fish swarm event timestamps and compares them against the video filenames. Each video file is named following a convention that includes

the date and time of recording, down to the second. The algorithm iterates through the list of video filenames within the directory for the given date and extracts the timestamps embedded in these filenames. Videos with the smallest temporal discrepancy to the event timestamp are selected, ensuring a close correspondence between observed behaviour and recorded video. Due to variations in the exact seconds at which videos start and the moments when fish behaviours are noted, an exact match between observation timestamps and video file names is improbable. The process that simplifies the matching protocol by reducing the resolution of the timestamp to a level that corresponds with the labeling of the video files. This time rounding respects the periodicity of the recordings and the observed events, ensuring that the subsequent comparison operates on a uniform timescale (Table[3]).

Table 3. Alignment of fish swarm event timestamps with corresponding video filenames, demonstrating the efficacy of the temporal matching algorithm.

Timestamp Rounded	**Species**	**Count**	**Video Path**	**Closest Video Datetime**	**Time Diff (min)**
2019-03-25 01:04:00	Roach	7	[20190325/010354856.avi, 20190325/010928869.avi]	2019-03-25 01:03:54	0.10
2019-03-25 04:07:00	Roach	5	[20190325/040640194.avi, 20190325/041207014.avi]	2019-03-25 04:06:40	0.33
2019-03-25 04:20:00	Roach	7	[20190325/041954734.avi, 20190325/042529308.avi]	2019-03-25 04:19:54	0.10
2019-03-26 21:30:00	Roach	5	20190326/213031267.avi	2019-03-26 21:30:31	0.52
2019-03-28 00:32:00	Roach	5	20190328/003205234.avi	2019-03-28 00:32:05	0.08

Outcome of the Matching Process - the outcome of the Swarm Identification and Video Matching stage is a dataset where each entry signifies a potential swarm event, now linked with the most temporally proximate video file (Fig[14]). This dataset is now augmented, pairing observational data with a visual record that precisely represents the behaviour of interest at or near the time it was noted. This matching process is essential in studies where behaviour must be corroborated visually. It allows for subsequent qualitative analyses of the fish swarm behaviours and supports quantitative assessments such as the frequency, duration, and composition of swarms.

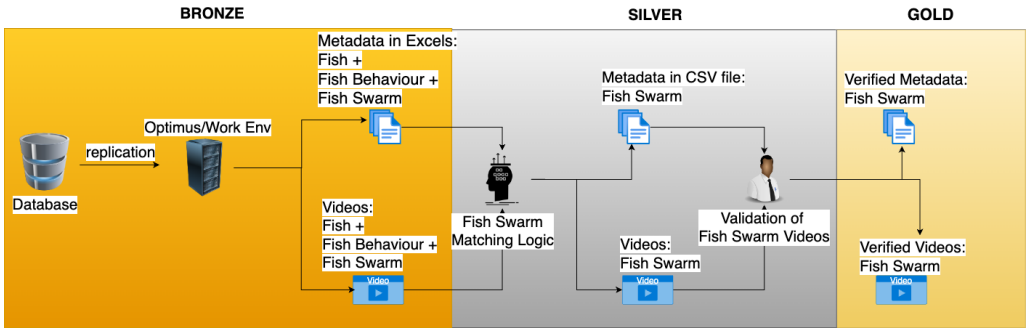


Figure 14. Data processing flow of 2019 FISH SWARM/Koblenz dataset.

Manual Validation Process - following the temporal matching and algorithmic selection of video files, each candidate video identified as potentially depicting a swarm event underwent a manual review. The manual validation process serves multiple purposes:

1. **Confirmation of Swarm behaviour** - each video is evaluated to determine whether it truly captures the dynamism and characteristics of a swarm, as opposed to isolated or unrelated fish movements.
2. **Integrity Assurance** - the manual review acts as a quality control measure, safeguarding against false positives that may have arisen during the algorithmic filtering and matching process.
3. **Refinement of Algorithmic Accuracy** - the insights gained from manual validation feed back into the system, honing the precision of the algorithms used for initial swarm event prediction.

The data processing for the 2020 fish swarms dataset, as depicted in the attached architectural flow diagram (Fig[15), exemplifies a streamlined and refined approach, primarily because the dataset had already undergone a level of pre-validation by human raters. Unlike the 2019 dataset, which required extensive preprocessing, the 2020 dataset was poised for a more direct application in the analytical model.

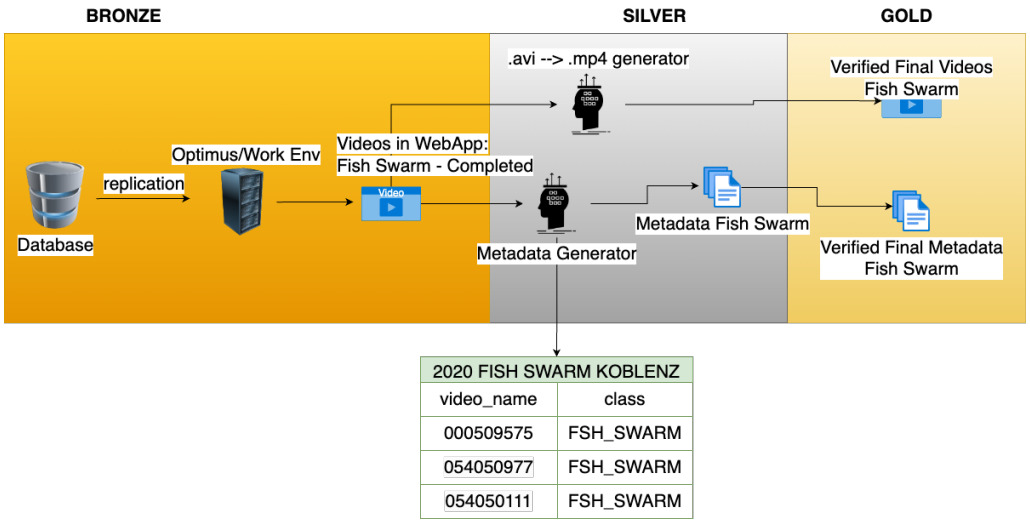


Figure 15. Data processing flow of 2020 FISH SWARM/Koblenz dataset.

Initial Data Replication - the process commences with the replication of the 2020 dataset into the "Optimus/Work Env" environment. This replication is a crucial step as it ensures that the processing and analysis are conducted within a controlled and resource-optimized setting. Replication also serves to safeguard the original data integrity while enabling multiple iterations of data handling without the risk of data corruption or loss.

Metadata Generation and Validation - once the dataset is securely integrated into the working environment, the next phase is metadata generation. For the 2020 dataset, the absence of accompanying metadata implies a need to generate descriptive information that provides context to the video content. This metadata typically includes timestamps, fish video class parameters that are essential for accurate data classification and retrieval.

The metadata generation and file conversion process is conducted using python scripts designed to parse video filenames and internal timestamps, generating structured metadata that aligns with the data model used in the analytical processes.

3.1.2 Fish & Fish Behaviour

The processing of Fish and Fish behaviour videos from 2020–2021, collected from Site-1 and Site-3, constitutes a significant phase in the data preprocessing architecture (Fig[16). This phase begins with the critical step of replicating the collected video data onto the 'Optimus' server. This server acts as the central working and development environment, ensuring that the data remains accessible and secure for processing. The replication process is meticulous, involving the transfer and verification of data integrity to ensure that no video data is corrupted or lost during the transition.

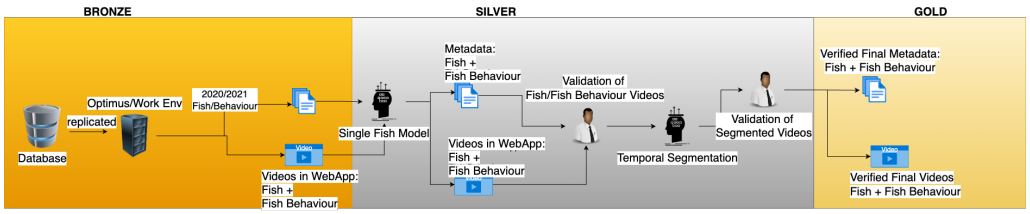


Figure 16. Data processing flow of 2020/2021 FISH and FISH BEHAVIOUR dataset.

After successful replication, the next step involves the classification of these videos into single and multiple fish categories using the established Single-Fish Model (SFM). This model, integral to the preprocessing of the video data, operates with a 79% accuracy rate. While this level of accuracy provides a substantial reduction in dataset complexity by filtering out videos unlikely to contain single fish, it is insufficient for complete automation of the process. Consequently, videos classified as containing single fish undergo a rigorous manual validation process to ensure the precision of the model's predictions.

Manual validation is a labor-intensive and critical component of the dataset preparation, involving several detailed steps:

1. Individual Video Review: Each video predicted as containing a single fish is meticulously watched by trained researchers. This step is crucial to confirm the presence of single fish and to assess the video's suitability for further analysis based on predefined research criteria.
2. Metadata Preparation: For each video undergoing manual validation, researchers prepare a comprehensive metadata file. This file captures essential details such as the video's unique identifier, recording site, and its classification as either a Fish or Fish behaviour type video. It also includes binary indicators for single fish presence (is_single), validation status (verified), and timestamps (start_1, end_1, start_2, end_2) that specify intervals of observed single fish activity.
3. Video Cropping: Utilizing the intervals specified in the metadata, an automated Python script crops the videos to focus exclusively on the segments of interest. This step is designed to isolate and enhance the study's focus on single fish behaviours, minimizing distractions from non-relevant content.
4. Second Manual Validation: After cropping, videos are subjected to another round of manual validation. This step is critical to ensuring that the video segments have been correctly identified and extracted according to the specified intervals. It also reassesses the categorization of each video, confirming its classification as a single fish video.
5. Final Dataset Compilation: Post-validation, the refined dataset of fish and fish

behaviour videos from Sites 1 and 3 is amalgamated with a similarly validated dataset of fish swarms from Koblenz. This final stage involves integrating various datasets to create a unified repository ready for complex analytical tasks.

The manual validation process demands substantial human effort and expertise, reflecting the project's commitment to data accuracy and reliability. Each video is scrutinized to ensure that the dataset upholds the high standards required for valid scientific inquiry. The detailed metadata, combined with precise video cropping and rigorous re-validation, underscores the meticulous nature of this research phase. It took around a week for a group of three people to validate the fish and fish behaviour videos, and the validation of cropped videos with given time intervals took another 4-5 days. This ensures that subsequent analyses on fish behaviour and dynamics are based on the most reliable and precise data available.

3.2 Fish - No Fish (FNF) Model

The Fish - No Fish (FNF) model employed in this research is based on the YOLOv5s architecture, as provided by Ultralytics [36]. This robust framework performs binary classification (fish or no fish) on underwater video footage, distinguishing sequences that contain fish from those that do not. The FNF model serves as a crucial first step for automated fish counting systems, which are increasingly used to monitor fish populations in diverse underwater environments.

The YOLOv5s model, known for its speed and efficiency, is well-suited for real-time applications. The architecture comprises three main components: the backbone, the neck, and the head. The backbone utilizes a modified CSPDarknet53 to extract essential features from the input image through convolutional layers and Cross Stage Partial (CSP) connections, which improve gradient flow and reduce model size. The neck includes Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PANet) modules that generate feature pyramids, enhancing the model's ability to detect objects at multiple scales. Finally, the head consists of YOLO layers that apply anchor boxes to the features and predict bounding boxes, objectness scores, and class probabilities for each detected object, covering different scales to accurately detect small, medium, and large objects.

The FNF model was trained using a dataset of approximately 77,000 manually annotated frames, ensuring high-quality training data. The model operates with an inference size of 640x640 pixels and a minimum confidence threshold of 0.5, configurable within the FNF script. This means that the model only detects bounding boxes with a confidence level above the threshold, ensuring reliable detection results.

Path from Dataset to Bounding Box Information - once the dataset is prepared and validated, incorporating meticulously matched video footage with corresponding metadata, the FNF model comes into play. It processes each video frame, applying binary classification to determine the presence of fish. Upon detecting fish, the model then employs bounding box (bbox) regression techniques to outline each individual fish within the frame (Fig[17). These bbox coordinates are instrumental in creating a feature space for further analysis. They enable precise tracking of fish movements and behaviours, providing rich data that feeds into machine learning algorithms for swarm behaviour detection.

The progression from raw video to bbox information involves the following steps:

- Preprocessing of Videos: Videos are preprocessed based on environmental conditions, enhancing image quality for more reliable fish detection.
- Application of the FNF Model: The FNF model processes the preprocessed frames, classifying them as containing a fish or not.
- Bounding Box Extraction: For frames classified with fish presence, the model applies bbox extraction to delineate each fish, generating spatial coordinates of the detected fish in the frame.
- Feature Space Construction: The bbox information, along with other derived attributes such as fish size and movement vectors, is used to construct a feature space. This feature space is then leveraged for detailed analysis and modeling of fish swarming behaviour (Section 3.3.2).

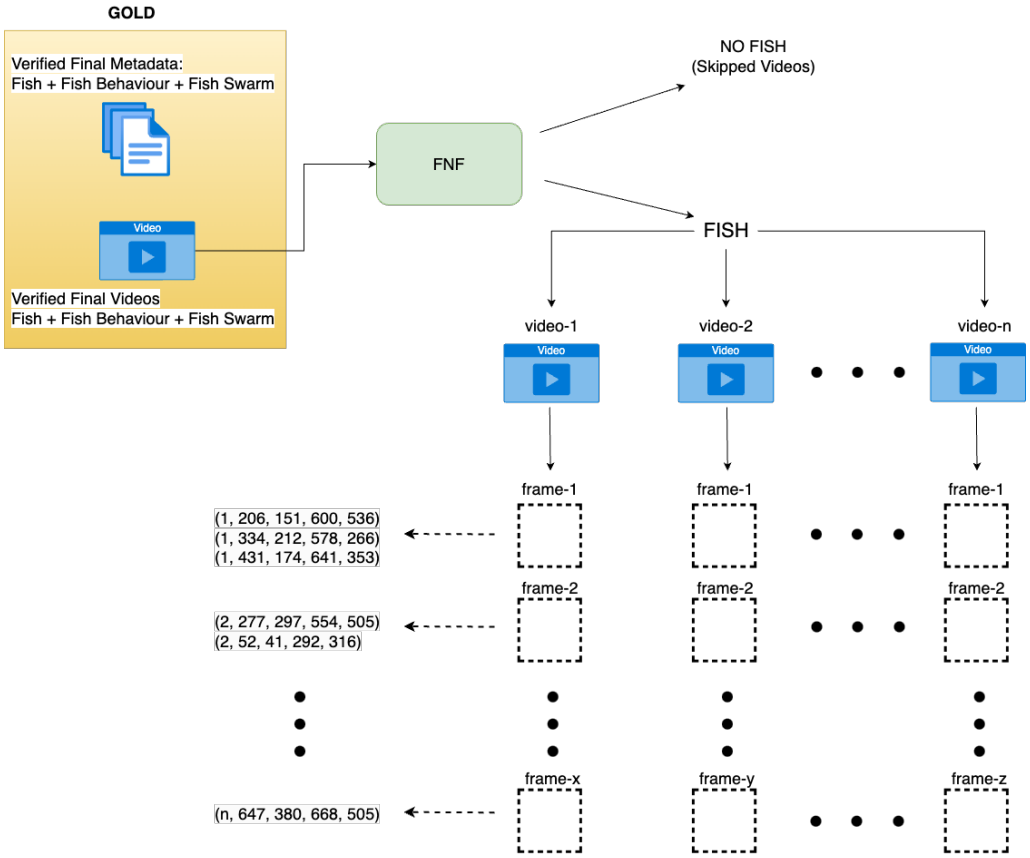


Figure 17. Bbox information retrieval process from each frame of the underwater fish video passing through FNF model.

3.3 Gradient Boosting Classifier

3.3.1 Model Overview

The choice of the Gradient Boosting Classifier (GBC) for this study was underpinned by a detailed comparative analysis of its performance against other well-regarded classifiers in tasks akin to fish swarm detection. The rationale for opting for GBC over other classifiers, such as Decision Tree, Random Forest, and Extra Trees, revolves around several key factors that align with the specific demands of the dataset and the objectives of this research.

The decision to utilize GBC was primarily influenced by its superior accuracy noted in previous studies, where it achieved up to 88.9% accuracy (Table[2]). This is notably high, especially considering that this performance was attained without the application of data scaling techniques, which are often necessary to enhance the performance of other classifiers. GBC's ability to handle raw, untransformed data effectively is particularly valuable in ecological datasets where preprocessing can sometimes distort critical natural

variances.

Table 4. A result of Classifiers' performance across different scalers.

Classifier	**Decision Tree**	**Random Forest**	**Extra Tree**	**Gradient Boost**
No scaler	86%	88%	87.1%	**88.9%**
No scaler simple	85%	84%	82.5%	85.9%
Standart scaler	86.1%	87.7%	87.1%	**88.9%**
Standard scaler simple	88%	84.8%	85.4%	88%
Robust scaler	86%	88.3%	87.1%	**88.9%**
Robust scaler simple	88%	85.1%	85.4%	88%

This advanced machine learning algorithm is known for its predictive accuracy, particularly in complex datasets where relationships between variables are non-linear. GBC operates by building an ensemble of weak prediction models, typically decision trees, that collectively form a robust predictive model. In the referenced study, the classifier was tested across various configurations and preprocessing scales, consistently maintaining high performance. This ability to handle diverse and relatively small datasets, such as those encountered in fish swarm detection where the behaviour of swarms versus solitary fish presents unique challenges, made it a compelling choice for our current research.

Rationale for Classifier Selection In the referenced study, the robustness of GBC was evident as it consistently outperformed other models across different feature sets and scaler conditions. Notably, the simple model configurations (using a single depth or estimator parameter) also yielded high accuracy, which underscores the efficiency of GBC in leveraging complex heuristical features for classification tasks. These features included:

1. Number of object detections throughout the video at varying confidence thresholds.
2. Maximum number of detections in a single frame, also at varying confidence levels.
3. Number of tracks and the average and maximum confidence per track.
4. Maximum area and density of bounding boxes within the video frames.

Given the similarity in the type of data and the nature of the classification problem in our current study, these results significantly influenced the decision to employ GBC for detecting fish swarms. The classifier's ability to perform well with a relatively small dataset (only 379 swarm and 379 non-swarm videos) and its resilience against overtraining are particularly advantageous for our research context, where video data are complex and labeling is labor-intensive.

The Gradient Boosting Algorithm (GBA) was utilized as the primary method for predictive modeling in this study provided by the python scikit-learn library. This method is

particularly adept at navigating complex data landscapes where traditional linear models falter due to non-linear relationships and interactions among variables [37]. The specific architecture and configuration used in this research include the following components and hyperparameters:

1. Base Estimators:
   - Decision Trees: The primary building blocks of the GBC are decision trees, specifically regression trees for classification tasks. Each tree is a weak learner that focuses on correcting the errors made by the previous trees in the ensemble.
2. Ensemble Method:
   - Additive Model: The GBC builds an additive model in a stage-wise fashion. This means that the model is built sequentially by adding one tree at a time, and each new tree corrects the errors of the combined ensemble of all previous trees.
3. Loss Function:
   - Deviance (Logarithmic Loss): For classification tasks, the default loss function is deviance, also known as logistic loss or binomial deviance. This loss function measures the difference between the predicted probability and the actual class label, providing a measure of model accuracy that the algorithm seeks to minimize.
4. Boosting Process:
   - Gradient Descent: The GBC uses gradient descent to minimize the loss function. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions and fits a new tree to the negative gradient (residuals). This tree is then added to the ensemble with a weight determined by the learning rate.
   - n_estimators: Specifies the number of boosting stages or trees in the ensemble. Each boosting stage attempts to correct the errors of the previous ensemble.
   - Learning Rate: The learning rate parameter shrinks the contribution of each tree to the final model, controlling the step size in the gradient descent process. Lower learning rates lead to more robust models but require more trees to achieve the same performance.
5. Regularization Techniques:
   - Maximum Depth of Trees (max_depth): Limits the depth of each tree to prevent overfitting by controlling the model's complexity.
6. Randomization:
   - Controls the randomness of the estimator. It affects the sampling of the training data and the shuffling of the data to ensure reproducibility.

In this study, GBC is used to classify video data into fish swarm and non-swarm activities based on a feature vector derived from the FNF model's bounding boxes (Fig[18). These features capture various aspects of video frames, such as detection counts, confidence levels, bounding box density, and more, reflecting the dynamic and varied nature of aquatic environments (see Section 3.3.2). The classifier is trained using a balanced dataset comprising equal numbers of swarm and non-swarm videos to ensure unbiased learning and generalizability.

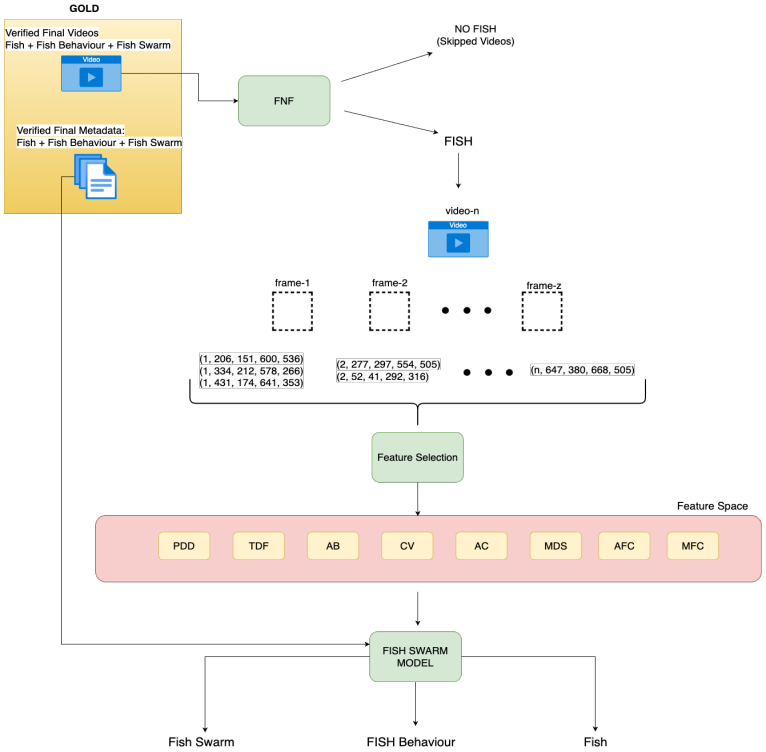


Figure 18. Design Architecture for Fish Behaviour Video Classification, final validated dataset of underwater fish videos in Gold layer are passed through the FNF model and retrieved bbox information of each fish detected frame, feature vectors are built on top of the bbox information and provided as an input for Fish Swarm model.

At each stage of the algorithm, the loss function, which measures the difference between the actual and predicted values, is calculated. The gradient of this loss function with respect to the predictions is then computed. This gradient information guides how the model's predictions should be adjusted to minimize the loss.

A new weak learner is introduced at each iteration to predict the residuals or errors of the ensemble thus far. These learners are typically shallow decision trees. By focusing on correcting the mistakes of previous learners, the algorithm iteratively improves the model's accuracy.

The output of each weak learner is scaled by a factor known as the learning rate before it is added to the ensemble. The learning rate, a value between 0 and 1, controls the speed at which the algorithm learns, acting as a form of regularization to prevent overfitting by making the model more robust to noise in the training data.

To ensure robust evaluation and to mitigate any potential biases in the model, 10-fold cross-validation is employed. This method partitions the data into ten subsets, facilitating the iterative training and validation of the model across all data subsets. This approach not only aids in assessing model performance across diverse data samples but also enhances the reliability and validity of the classifier by providing a comprehensive view of its predictive capabilities under different conditions.

3.3.2 Feature Space

The designed feature space for the fish swarm detection model is a multifaceted construct, aiming to capture the intricacies of underwater activity and fish behaviour. Here the features are listed and explained:

- **Total Detections per Frame** - TDF represents the count of bounding boxes detected in each frame, providing an immediate sense of activity within the video. High detection counts can indicate dense fish populations or swarming behaviour, especially in frames where multiple entities are detected simultaneously. This metric is essential for initial filtering of active versus inactive video segments and for setting the context for more detailed analysis in subsequent frames.
- **Average Confidence** - AC derived from the detection confidence scores of the bounding boxes, the average confidence metric reflects the model's certainty in its detections. Higher confidence scores are generally correlated with clearer and more distinguishable features within the bounding boxes, which are crucial for reliable classification in challenging underwater environments. This feature helps in weighting detections by reliability, prioritizing high-confidence detections in the classification process.
- **Minimum Detection Size** - MDS captures the smallest bounding box area detected in a video, providing insights into the minimum visible size of detected objects. In the context of fish detection, smaller sizes might indicate distant or smaller fish, which are important for assessing the range of fish sizes within a swarm. This metric is vital for differentiating between environmental debris and small fish, a common challenge in underwater video analysis.
- **Peak Detection Density** - PDD measures the highest number of detections within a single frame or a sequence of frames. It is indicative of peak activity periods, such

as when a swarm might be passing through the frame. High peak densities can be a strong indicator of swarming behaviour, making this feature particularly critical for distinguishing fish swarms from solitary or sparsely distributed fish.

- **Average Frame Coverage** - AFC calculates the mean proportion of the frame area covered by detections, offering a sense of how much of the video frame is occupied by detected objects over time. It is useful for understanding the spatial distribution of fish within the frame, with higher coverage typically associated with denser aggregations of fish, potentially indicating swarming behaviour.
- **Maximum Frame Coverage** - Similar to average frame coverage, MFC provides a snapshot of the maximum extent to which detections fill the frame. This feature can be crucial in identifying frames where the activity reaches its peak, possibly highlighting key moments of interaction or significant movement within the fish population.
- **Centroid Variance (X and Y)** - CV represents the variance in the x and y coordinates of the centroids of detections, offer insights into the dispersion and movement patterns of the detected objects across frames. High variance might suggest widespread activity across the video frame, while low variance could indicate stable, centralized swarming activity. These features help in understanding the dynamics of movement within detected groups, crucial for distinguishing between random movement and coordinated swarming.
- **Activity Bursts** - AB quantifies the number of frames with detection counts significantly above the defined threshold value which is set as five (Section 2.1.3), that may indicate sudden spikes in activity. This feature is developed in this study as a critical marker of fish swarm behavioural patterns that exhibit reactive or interactive dynamics within groups.

Each of these features plays a specific role in the model's ability to effectively classify underwater footage into fish swarm and non-fish swarm scenarios. By capturing a range of spatial and temporal characteristics from the video, these features allow the Gradient Boosting Classifier to make informed predictions based on comprehensive and robust data inputs, tailored to the unique challenges of underwater video analysis.

3.3.3 Evaluation

To effectively gauge the performance of the Gradient Boosting Classifier (GBC) used for detecting fish swarms in underwater video footage, we employ several key statistical metrics: accuracy, precision, recall (sensitivity), and the F1 score. These metrics provide insights into the accuracy and efficiency of the classifier relative to a manually annotated ground truth. Videos classified as FISH_SWARM or non-swarms (FISH or FISH_BEHAVIOUR)

were compared to human labels which served as the ground truth. If a video was classified as FISH_SWARM and the ground truth was also FISH_SWARM, this was considered as a True Positive (TP). If the classified video and the ground truth were both classified as NON_SWARM (FISH or FISH_BEHAVIOUR), the automated classification represent a True Negative (TN). If a video was classified as FISH_SWARM while the ground truth was non-swarms (FISH or FISH_BEHAVIOUR), this was counted as a False Positive (FP). Lastly, if the video classified as non-swarms (FISH or FISH_BEHAVIOUR) which should have been classified as FISH_SWARM, then it was assigned as False Negatives (FN).

These classifications form the foundation for our evaluation metrics, which are calculated as follows:

Accuracy - this metric evaluates the overall effectiveness of the classifier by measuring the proportion of total correct predictions (both true positives and true negatives) made out of all predictions. Accuracy is particularly useful for providing a quick snapshot of the model's performance across all classes, both positive and negative. It is calculated with the formula:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Observations}}$$

Precision - measures the accuracy of the fish swarm detections by the model, indicating the proportion of correct positive identifications out of all positive identifications made by the model. It is calculated using the formula:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall (Sensitivity) - Assesses the model's ability to identify all relevant instances of fish swarms from the dataset. This metric is crucial for understanding how effectively the model can detect swarms without missing any. The recall is computed as:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

F1 Score - The harmonic mean of precision and recall, the F1 score provides a balanced measure that considers both the precision and the recall of the classifier. This is particularly

important in scenarios where an equal weight is needed for both false positives and false negatives. It is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are derived from a 10-fold cross-validation method, which is employed to rigorously evaluate the GBC. Cross-validation ensures that every data point gets to be tested, which is crucial in small or imbalanced datasets like ours. This method also helps in mitigating overfitting and provides a more generalized performance indicator across different subsets of data.

4. Results

4.1 Dataset Preparation

In the 2019 fish swarm dataset from Koblenz, a comprehensive preprocessing initiative was undertaken on a corpus of 14,523 videos, which included various classifications such as fish, fish swarm, and fish behaviour. The initial dataset provided below analysis for the richness of the 2019 dataset:

Overall Fish Counts by Species:

- Small Cyprinids: 7,591 occurrences
- Perch: 4,579 occurrences
- Bleak: 4,180 occurrences
- Roach: 3,472 occurrences

These species dominate the observations, indicating their prevalence and possibly gregarious nature in the observed environment (see Fig[19]).

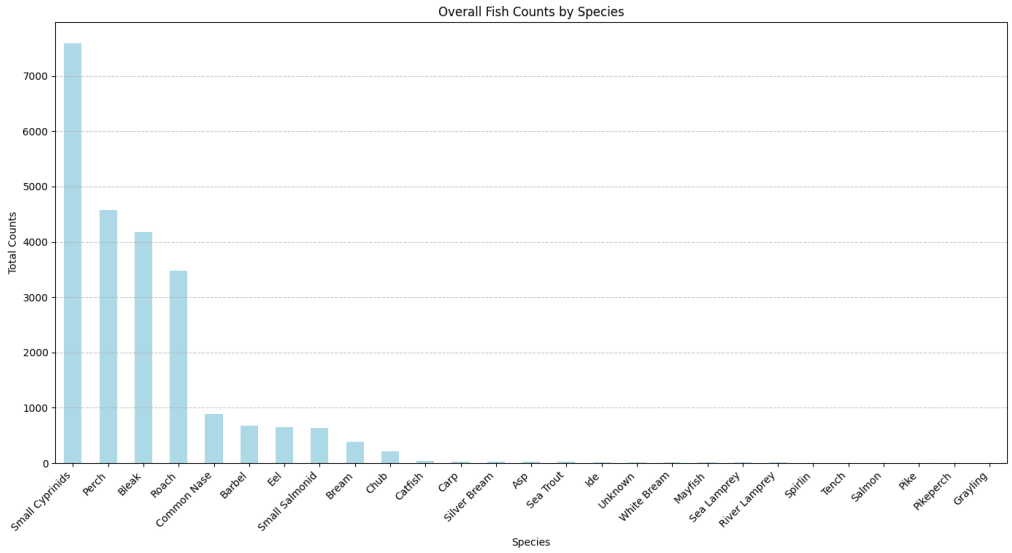


Figure 19. Overall Fish Counts by Species.

Total fish counts by time of day: the hourly distribution of fish counts shows peaks at (see Fig[19]):

- 4 AM: 1,414 counts
- 11 AM: 1,515 counts
- 5 PM: 1,889 counts
- 6 PM: 1,878 counts

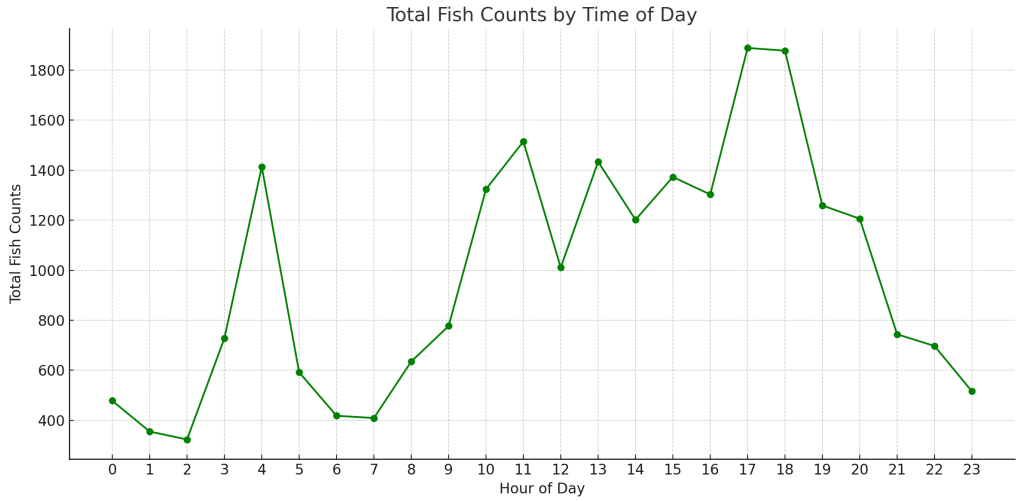


Figure 20. Total fish counts by time of day.

Upon applying the Fish Swarm Generator Logic to this raw data, a subset of 306 videos was initially identified as depicting fish swarm activities. Through rigorous manual validation processes, this number was refined to 297 videos, affirming the high precision of the classification logic with a 97% accuracy rate.

In subsequent analyses for 2020, the fish swarm video dataset expanded to include 745 new entries, culminating in a total of 1,042 videos when combined with the validated 2019 entries. This dataset was enriched with comprehensive metadata, enhancing its utility for further ecological and behavioural studies.

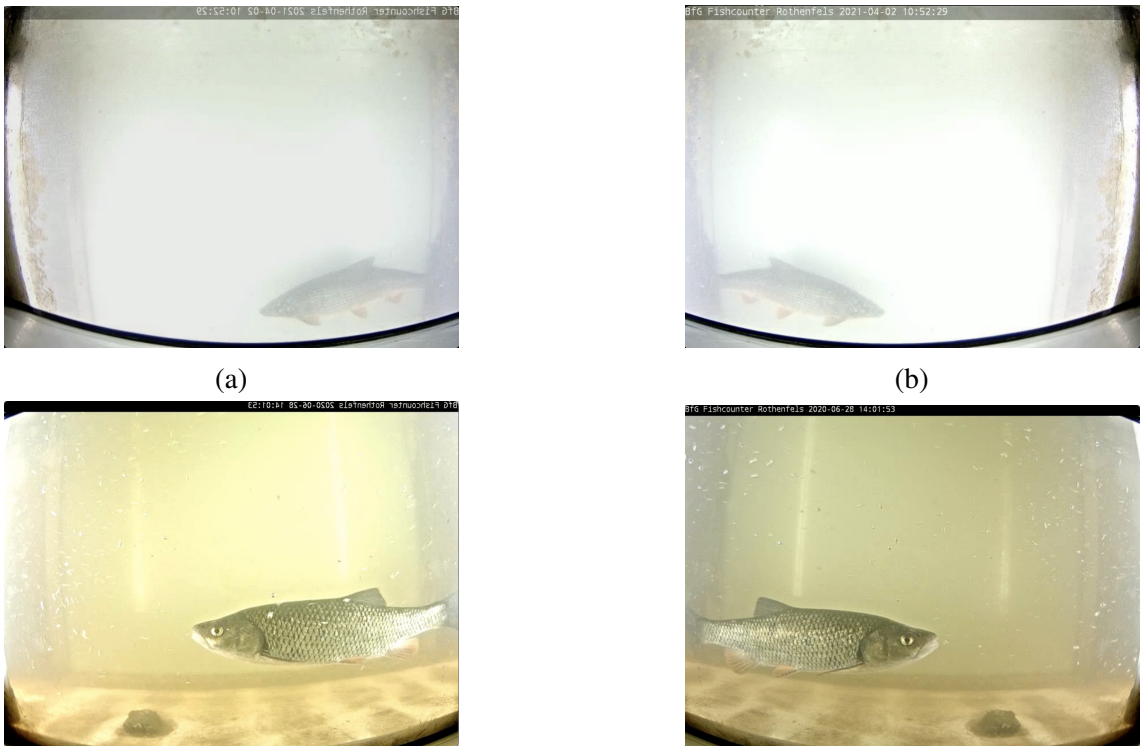
Parallel efforts focused on classifying fish and fish behaviour videos from site-1 and site-3, initially yielding 1,507 videos. Intensive manual validation, including detailed video segmentation, meticulously reduced this figure to 1,153 videos. This reduction was primarily distributed as 174 videos from site-1 and 140 from site-3 for fish type videos, alongside 220 and 619 videos for fish behaviour from sites 1 and 3, respectively (Table[5]).

Table 5. Total Number of Videos before Augmentation.

	site-1	**site-3**	**site-4**	**Total**
FISH	174	140	0	314
FISH BEHAVIOUR	220	619	0	839
FISH SWARM	0	0	1042	1042
Total	394	759	1042	2195

The skewed distribution towards fish and fish swarm/behaviour videos highlighted a significant imbalance within the dataset. To rectify this disparity and ensure a balanced representation across categories, augmentation techniques were employed for fish class videos.

The augmentation involved adjusting the brightness and contrast of the video frames to simulate varying environmental conditions that could affect visibility underwater. The script adjusts the brightness of each frame towards a target average level, carefully controlling the scaling to prevent excessive image degradation. For instance, the brightness adjustment factor is limited between 0.5 and 2 times the original brightness to maintain natural visual quality. Additionally, a slight contrast enhancement is applied based on the brightness adjustment, enhancing the visibility and differentiation of features within each frame (Fig.21).



(a) (b)

(c) (d)

Figure 21. Comparison of Original and Augmented Fish Class Videos from Multiple Sites. Panels (a, c) show videos augmented with horizontal flips and adjustments to brightness and contrast (b, d).

These adjustments were applied systematically across videos in the 'Fish' category to generate augmented versions. Each video was processed to flip the frames horizontally, simulating different orientations of fish movement, and adjusting brightness to a standardized target, fostering a more comprehensive range of visual data. This process was managed through a Python script utilizing the imageio library for reading and writing video

frames, ensuring each frame was modified consistently and stored in a new augmented video file. The final number of videos for each class after augmentation is described in Table[6]).

Table 6. Total Number of Videos after Augmentation.

	site-1	**site-3**	**site-4**	**Total**
FISH	348	280	0	628
FISH BEHAVIOUR	220	619	0	839
FISH SWARM	0	0	1042	1042
Total	568	899	1042	2509

4.2 FNF Model

The Fish-No Fish model serves a pivotal role in the preprocessing of video data for this study. The model's primary function is to differentiate video frames containing fish from those without, providing essential bounding box (bbox) information for each detected fish. This classification is critical for filtering out non-relevant video footage, thereby focusing resources on frames that potentially contribute to understanding fish behaviour and swarming patterns.

Throughout the evaluation, the FNF model was applied to a substantial number of videos across various categories, where it efficiently identified frames with fish presence. This process not only ensures that the subsequent analysis focuses on relevant data but also enhances the efficiency of the computational pipeline by reducing the volume of data to be processed.

Key outcomes from the model's application are summarized as follows:

1. Total Videos Processed: The model reviewed a comprehensive set of 2509 videos from different behavioural categories.
2. Effective Detections: The FNF model successfully detected fish presence in 2108 videos, representing a significant portion of the total videos analyzed. These detections were instrumental for the next stages of feature vector creation and further analysis.

The videos processed through the FNF model that resulted in effective detections were further categorized into distinct classes based on the labels from the augmented datasets:

1. Fish Swarm Videos: Detected in 662 videos.
2. Fish Videos: Detected in 628 videos, including augmented versions aimed at balanc-

ing dataset representation.

3. Fish behaviour Videos: Detected in 818 videos.

These detections formed the basis for constructing a detailed feature vector dataset. Each detected frame provided bounding box coordinates which encapsulate the location and size of each fish within the frame. These metrics are crucial for the subsequent analytical steps that involve more refined behavioural analysis, including swarming behaviour.

A comprehensive correlation analysis is conducted to understand the relationships between the various features used in fish swarm detection model. Given the complexity of behaviours and environmental variables captured in the dataset, identifying these relationships aids in refining the feature engineering and selection processes.

A balanced dataset is utilized, consisting of three distinct classes: FISH, FISH_BEHAVIOUR, and FISH_SWARM. The dataset was balanced using resampling techniques to ensure equal representation of each class, thus preventing any class imbalance from skewing our analysis.

4.3 Feature Space

In our comprehensive analysis of the correlations between various features and the class label in whole dataset encompassing characteristics of fish, fish behaviour, and fish swarms, several significant relationships were revealed that are pivotal for enhancing our understanding of the data and refining our predictive models (see Fig[22]). Notably, total_detections and activity_bursts demonstrated a remarkably strong correlation (0.86), underscoring a potential intrinsic link between the frequency of detections and the occurrences of burst activities in underwater environments. This relationship suggests that regions with frequent fish detections are also areas of high activity, possibly indicative of swarm behaviours.

Further examination revealed a substantial positive correlation (0.56) between peak_detection_density and the class label, highlighting its utility in distinguishing between different classes effectively, particularly in identifying swarm behaviours versus solitary or less dense fish activities. This feature, when combined with centroid_variance_x which also showed a notable correlation with the class label (0.65), could provide robust indicators for classifying fish behaviours based on movement patterns and spatial distribution within the observed frame.

The correlation between average_frame_coverage and max_frame_coverage (0.74) was also significant, indicating that frames with higher average coverage tend to reach their

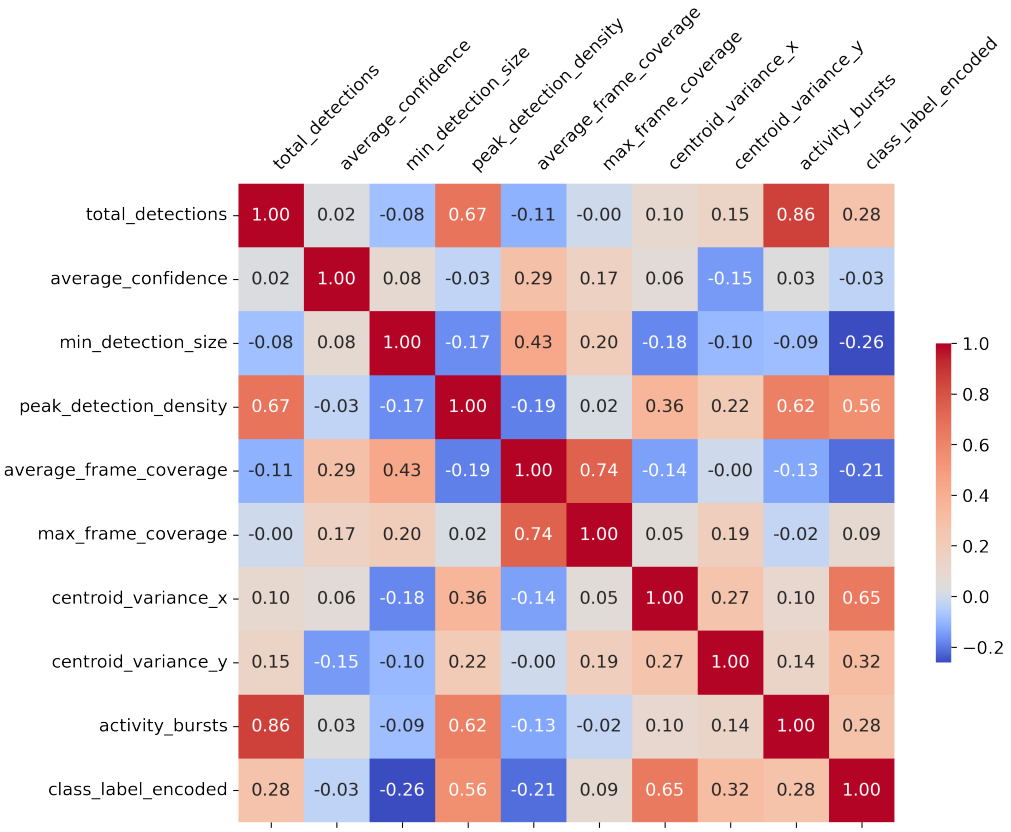


Figure 22. Feature Space correlation including Class label.

maximum coverage limits, which could be reflective of environmental or situational constraints affecting the fish populations. Intriguingly, the negative correlation between min_detection_size and the class label (-0.26) suggests that smaller detections are less likely to correlate with specific behaviours categorized under the current class labels.

The correlation analysis of the FISH_SWARM dataset reveals several key insights about the relationships between different features, which can help understand the behaviours and characteristics of fish swarms (see Fig[23]).

Key Insights from the Correlation Metrics:

1. Strong Positive Correlation Between Total Detections and Activity Bursts (0.841) - This strong correlation indicates that as the number of total detections increases, so does the number of activity bursts. This relationship suggests that higher fish activity, possibly an important factor in understanding swarm dynamics.
2. Moderate Positive Correlation Between Peak Detection Density and Activity Bursts (0.575) - The correlation between peak detection density and activity bursts suggests that more densely packed swarms are associated with higher levels of activity bursts. This might indicate that denser aggregations of fish are more active

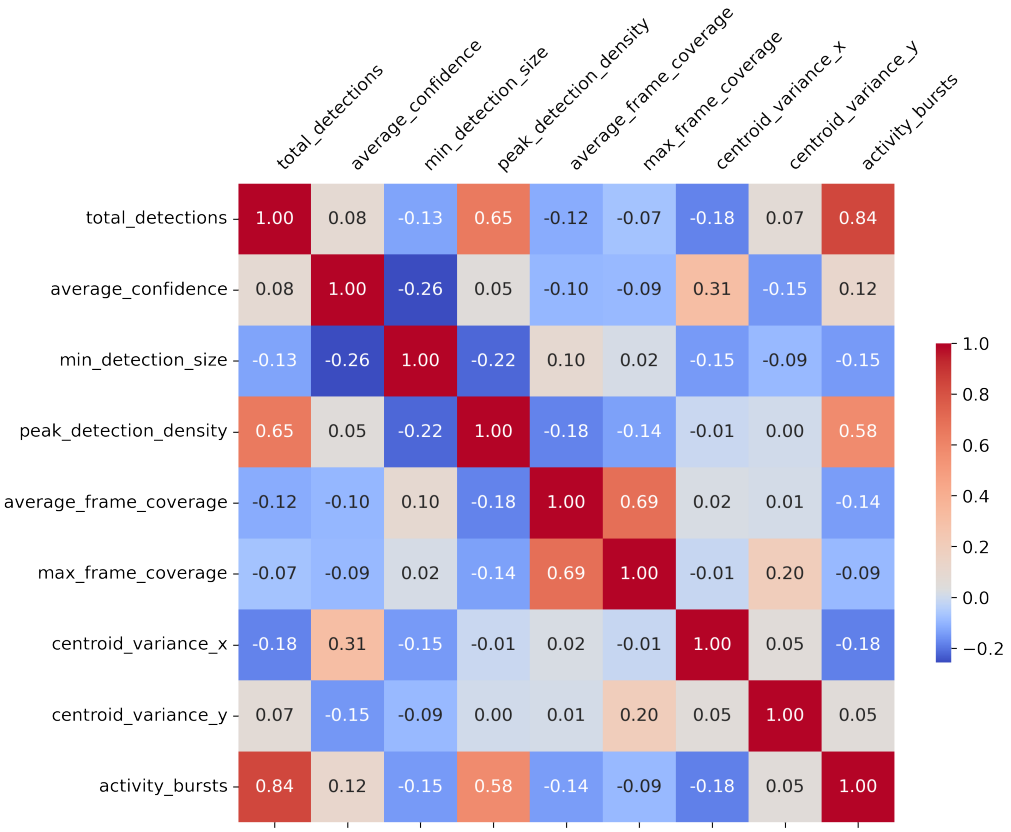


Figure 23. Feature Space correlation for FISH_SWARM Class.

3. Negative Correlations Involving Min Detection Size - Min detection size shows negative correlations with average confidence (-0.257), indicating that smaller objects tend to be detected with less confidence. This could suggest difficulties in accurately detecting smaller fish or smaller groups within a swarm, potentially due to resolution limitations or the behaviour of smaller fish.
4. High Correlation Between Average and Max Frame Coverage (0.691) - The strong positive correlation between average frame coverage and max frame coverage suggests consistency in how much of the frame is covered by detections. High frame coverage could indicate large swarm sizes or high density, which are crucial for understanding the extent and nature of fish swarms.
5. Negative Correlation Between Centroid Variance and Activity Bursts (-0.184) - The slight negative correlation between centroid variance and activity bursts could imply that more consistent positioning (lower variance in centroid positions) correlates with fewer bursts of activity, potentially indicating more stable or less agitated swarm states.

Observations indicate that increased detections and peak densities are significantly associated with more frequent activity bursts, underlining the potential of these metrics as indicators of swarm activity and density. Particularly, the strong positive correlation

between total detections and activity bursts suggests a direct relationship between the observed number of fish and their collective behaviours. Conversely, smaller detected sizes correspond with lower confidence levels, highlighting a possible area for technical improvement in detection algorithms.

Moreover, the correlation between average and maximum frame coverage provides insights into the spatial dynamics of swarms, with implications for understanding how environmental or interspecific interactions influence swarm formation and behaviour. Interestingly, the negative correlation between centroid variance and activity bursts could indicate a behavioural adaptation where swarms maintain a cohesive formation in less active states, possibly as a defensive mechanism against predators.

The correlation analysis for the FISH_BEHAVIOUR dataset (see Fig[24]), particularly emphasizing the dynamics of non-swarm activities, unveils several interesting patterns and relationships among the features. Given that the primary research interest does not focus on non-swarm activities, the insights derived here can still provide valuable context and serve as a baseline for understanding the more complex behaviours associated with fish swarms.

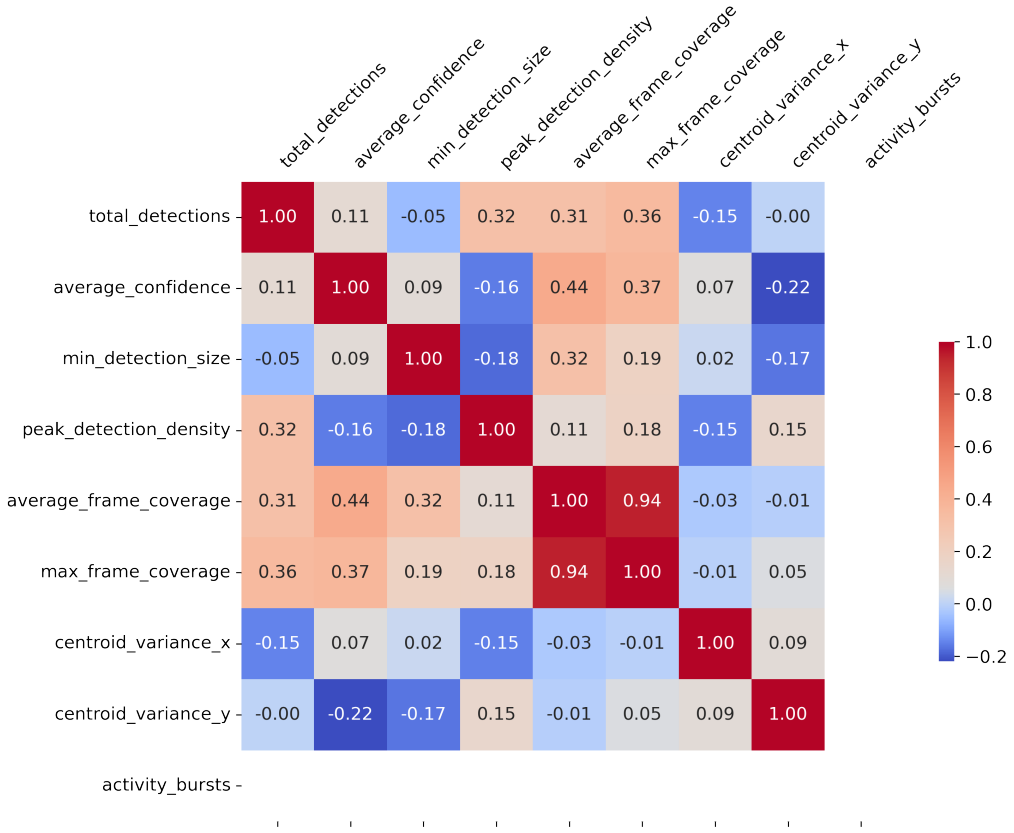


Figure 24. Feature Space correlation for FISH_BEHAVIOUR Class.

Key Insights from Correlation Metrics:

1. Moderate Positive Correlation Between Total Detections and Coverage Metrics - The correlations between total_detections and both average_frame_coverage (0.314) and max_frame_coverage (0.356) suggest that as the number of detections increases, there is a corresponding increase in the area of the frame that is covered. This could indicate that more active or numerous fish within the field of view contribute to higher coverage metrics.
2. Significant Positive Correlation Between Coverage Metrics - A very high correlation between average_frame_coverage and max_frame_coverage (0.944) indicates that these two metrics are almost redundant. This suggests that frames with higher average coverage tend to reach their maximum coverage consistently, potentially reflecting a uniform behaviour pattern across observations.
3. Negative Correlation Between Confidence and Detection Size with Certain Metrics:
    - The negative correlation between average_confidence and centroid_variance_y (-0.220) might suggest that lower confidence in detections is associated with greater variance in the y-coordinate of centroids, possibly indicating erratic behaviour or detection errors in less confident observations.
    - Similarly, the negative correlation between peak_detection_density and average_confidence (-0.157) could imply that denser groups are detected with slightly less confidence, possibly due to overlapping fishes or complex group dynamics.
4. Weak or Insignificant Correlations Involving Centroid Variances - Both centroid_variance_x and centroid_variance_y show minimal correlation with most other features, indicating that the variance in fish positions does not strongly influence other measured behaviours in non-swarm activities.
5. Non-Existent Correlations for Activity Bursts: - the NaN values for activity_bursts across all features highlight an absence of burst activities within the dataset, which aligns with the expectation that such behaviours are more typical of swarm classes rather than isolated or non-swarm behaviours.

The correlation analysis for the dataset classified under the FISH category (see Fig[25]), which represents non-swarm activities same as FISH_BEHAVIOUR, highlights several key interdependencies between the measured features. These findings can serve as a contrast to swarm behaviour, enhancing the understanding of varying behavioural dynamics across different underwater contexts.

Key Insights from Correlation Metrics:

- Positive Correlations Involving Detection Metrics and Coverage - Total Detections and Coverage Metrics: There is a moderate correlation between total_detections

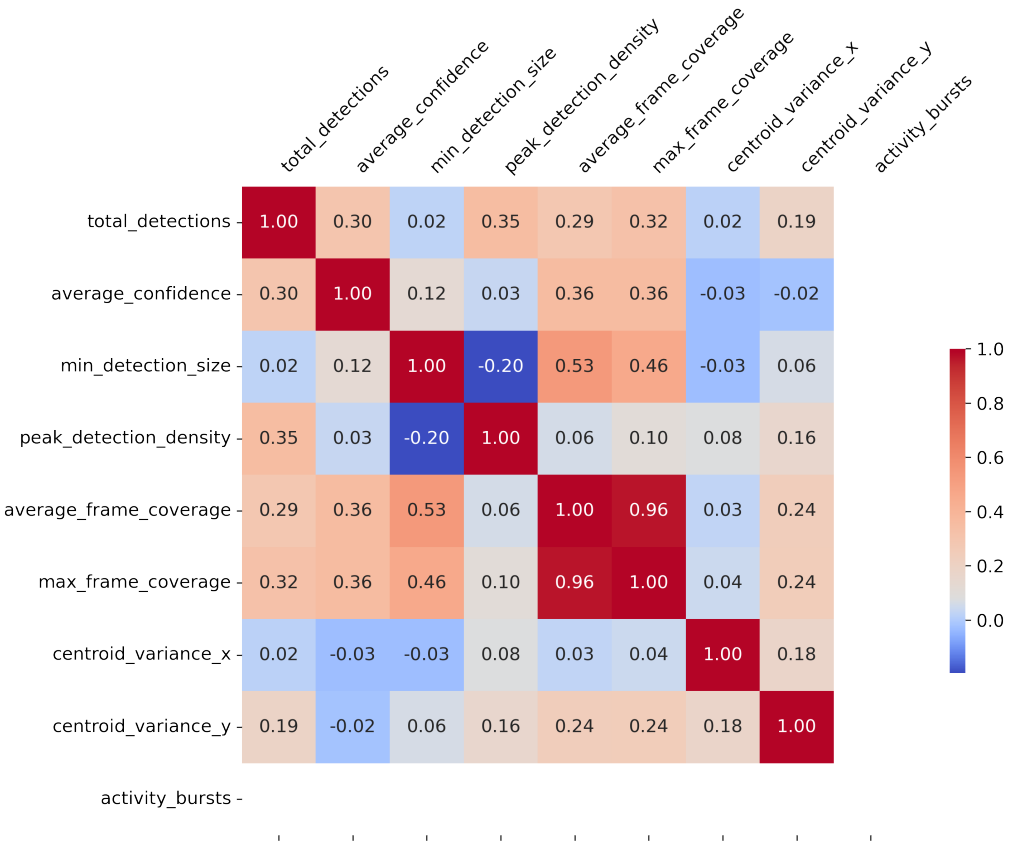


Figure 25. Feature Space correlation for FISH Class.

and both average_frame_coverage (0.290) and max_frame_coverage (0.325). This suggests that an increase in detections typically corresponds to a higher coverage of the observation area, likely indicating that more fish within the field of view increase the spatial extent of coverage.

- Average and Max Frame Coverage - The very high correlation between average_-frame_coverage and max_frame_coverage (0.965) indicates these two metrics are closely linked, with frames that have high average coverage generally reaching near their maximum potential coverage.
- Correlations Involving Min Detection Size - Min Detection Size and Coverage Metrics: The positive correlations of min_detection_size with average_frame_-coverage (0.533) and max_frame_coverage (0.461) imply that larger minimum detection sizes tend to be associated with greater frame coverage. This could reflect larger fish or more visually distinct fish being easier to detect and covering more area within the frame.
- Centroid Variance Correlations - Centroid Variance and Coverage: The positive correlations between centroid_variance_y with both average_frame_coverage (0.241) and max_frame_coverage (0.245) suggest that variations in the vertical distribution of fish within the frame slightly influence the coverage metrics. This might indicate vertical movements or depth variations among the fish being tracked.

- Lack of Activity Bursts - The NaN values for activity_bursts across all features highlight an absence of this particular type of dynamic behaviour in the dataset for the FISH class. This absence is in line with expectations as burst activities are typically associated with swarm dynamics rather than solitary fish behaviours.

4.4 Model Evaluation

The model's performance was rigorously evaluated using a combination of parameter tuning via GridSearchCV and cross-validation. a systematic approach was taken to select the most effective parameters for the GBC. This involved testing a range of values for key hyperparameters to determine the best combination for maximizing the accuracy of the model. Here's a detailed look at the parameters that were tested, the rationale for their selection, and the final parameters that were chosen as the best.

4.4.1 Hyperparameter Selection

To test the model with different hyperparameters, GridSearchCV is used from scikit-learn. Grid Search employs an exhaustive search strategy, systematically exploring various combinations of specified hyperparameters and their default values. This approach involves tuning parameters, such as learning rate, max depth, number of estimators, through a cross-validated model, which assesses performance across different parameter settings.

Table 7. A greed of hyperparameters used in cross-validation.

learning_rate	0.05	0.1	0.15
max_depth	3	4	5
n_estimators	50	75	100

After running the GridSearchCV, which performed cross-validation across different combinations of these parameters, the best-performing set of parameters were found to be:

- Learning Rate: 0.1
- Max Depth: 4
- Number of Estimators: 75

4.4.2 Dataset

The evaluation employed a balanced dataset to ensure unbiased performance assessment across all classes. Each class—FISH, FISH_BEHAVIOUR, and FISH_SWARM—was represented with 617 instances, totaling 1851 data points. This balanced approach is

crucial for avoiding model bias towards the more frequently represented classes.

4.4.3 Validation Results

The model's generalization capability was validated using a 10-fold StratifiedKFold technique. This method ensures each fold is representative of the overall dataset, maintaining the same proportion of each class across folds, thus providing a reliable estimate of the model's performance on unseen data.

Performance Metrics:

- Precision: The model achieved average precision of 0.87, with class-specific precision scores of 0.82 for FISH, 0.83 for FISH_BEHAVIOUR, and 0.98 for FISH_SWARM (see Table[8]).
- Recall: The average recall also stood at 0.87, with FISH achieving a recall of 0.87, FISH_BEHAVIOUR 0.80, and FISH_SWARM 0.94.
- F1 Score: The overall F1 score was 0.87, reflecting a balanced harmonic mean of precision and recall, with individual class scores closely mirroring the precision and recall results.

Table 8. Performance Metrics of Each Class after Cross-Validation.

	precision	**recall**	**f1-score**	**support**
FISH	0.80	0.85	0.82	617
FISH BEHAVIOUR	0.81	0.79	0.80	617
FISH SWARM	0.98	0.95	0.97	617
Accuracy			0.86	1851
Macro Avg	0.87	0.86	0.86	1851
Weighted Avg	0.87	0.86	0.86	1851

Confusion Matrix Analysis:

- 523 true positives for FISH with 94 instances misclassified as FISH_BEHAVIOUR and 0 as FISH_SWARM (Fig. [25]).
- 490 true positives for FISH_BEHAVIOUR, with misclassifications primarily as FISH (118 instances) and 9 as FISH_SWARM.
- 586 true positives for FISH_SWARM, underscoring the model's effectiveness in identifying swarm behaviours with only minor confusions with FISH_BEHAVIOUR (21 instances) and FISH (10 instances).

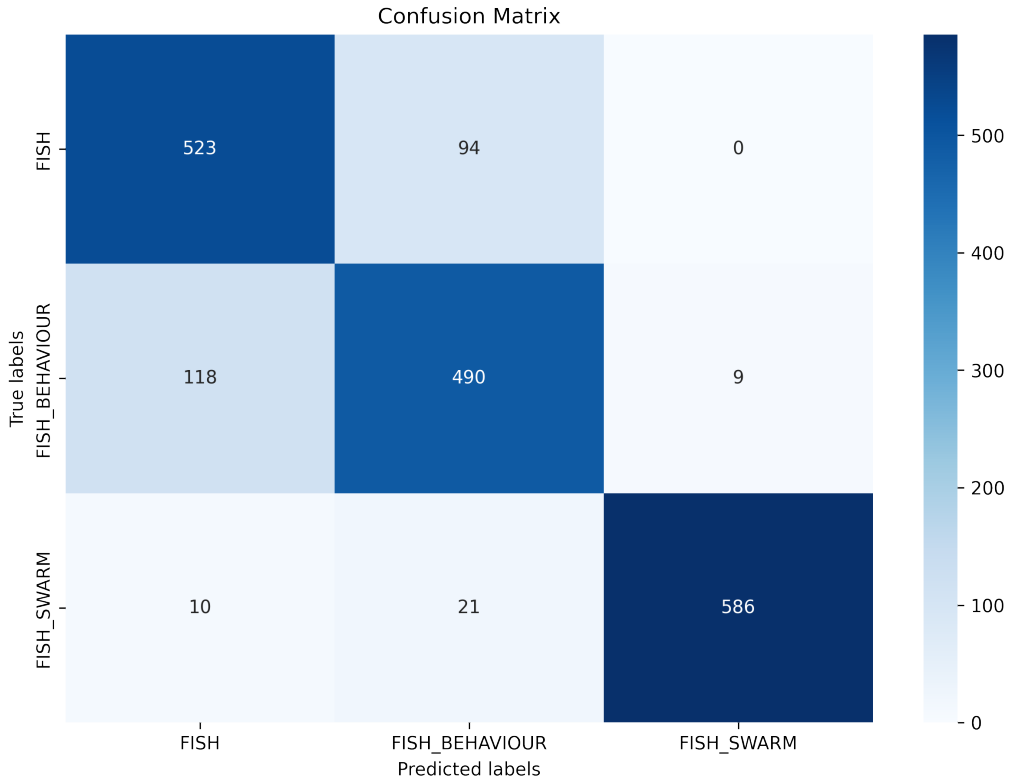


Figure 26. Confusion Matrics generated during the cross-validation phase of the designed Fish Swarm Model.

4.4.4 Training

Following the selection of hyperparameters through GridSearchCV, the model was trained using the optimal settings derived: an n_estimators value of 75, a learning rate of 0.1, and a max_depth of 4. The training and test datasets were split 80-20%. This configuration was aimed at achieving the best compromise between model complexity and performance while preventing overfitting.

The training process was monitored to evaluate how the model's performance evolved across iterations:

- Loss Reduction: During the training phase, the loss steadily decreased, indicating the model was effectively learning from the training data. The initial loss started at 0.9696 and was reduced to 0.1222 by the 75th iteration, reflecting consistent improvement in the model's ability to fit the data accurately.

The plotted deviance (Fig. 27) shows that the training and test sets start with a similar deviance close to 1. As the number of boosting iterations increases, the training deviance decreases rapidly, indicating that the model is learning and fitting the training data well. In

contrast, the test deviance decreases more slowly and eventually stabilizes. This widening gap between the training and test set deviance over iterations indicates that while the model continues to improve its performance on the training data, its performance on the unseen test data does not improve correspondingly after a certain point. Beyond approximately 50–70 iterations, the test deviance begins to stabilize, suggesting the model has reached its optimal capacity for learning from the data, and that further iterations are not providing significant gains. Therefore, training was stopped after 75 iterations to balance the model's performance and prevent overfitting.

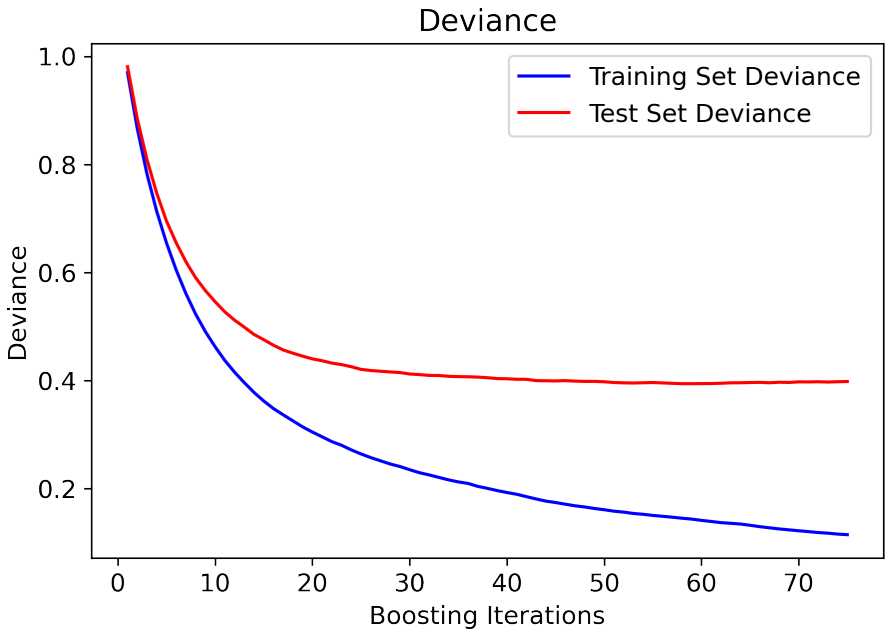


Figure 27. Training and Test Set Deviance Over Iterations for Gradient Boosting Model.

In Fig. 28, the key outcomes from the plots demonstrate the performance and behavior of the GBC over 75 boosting iterations. The recall plot shows a rapid increase in the early iterations, stabilizing around 0.525 after approximately 10-15 iterations, indicating that the model quickly learns to identify true positives effectively. Similarly, the accuracy plot exhibits a significant improvement in the initial iterations, plateauing around the same point, which reflects the model's ability to correctly classify both positive and negative instances. The precision plot, which measures the accuracy of positive predictions, initially rises sharply and then slightly decreases before stabilizing around 0.4, highlighting a brief period of increased false positives as the model aggressively improves recall. The F1-score, a harmonic mean of precision and recall, follows a pattern similar to recall and accuracy, with a rapid increase and subsequent stabilization around 0.425, demonstrating a balanced performance between precision and recall. The consistent stabilization of these metrics after 10-15 iterations suggests that the model reaches an optimal performance level early in the boosting process, with subsequent iterations providing minimal additional benefit.

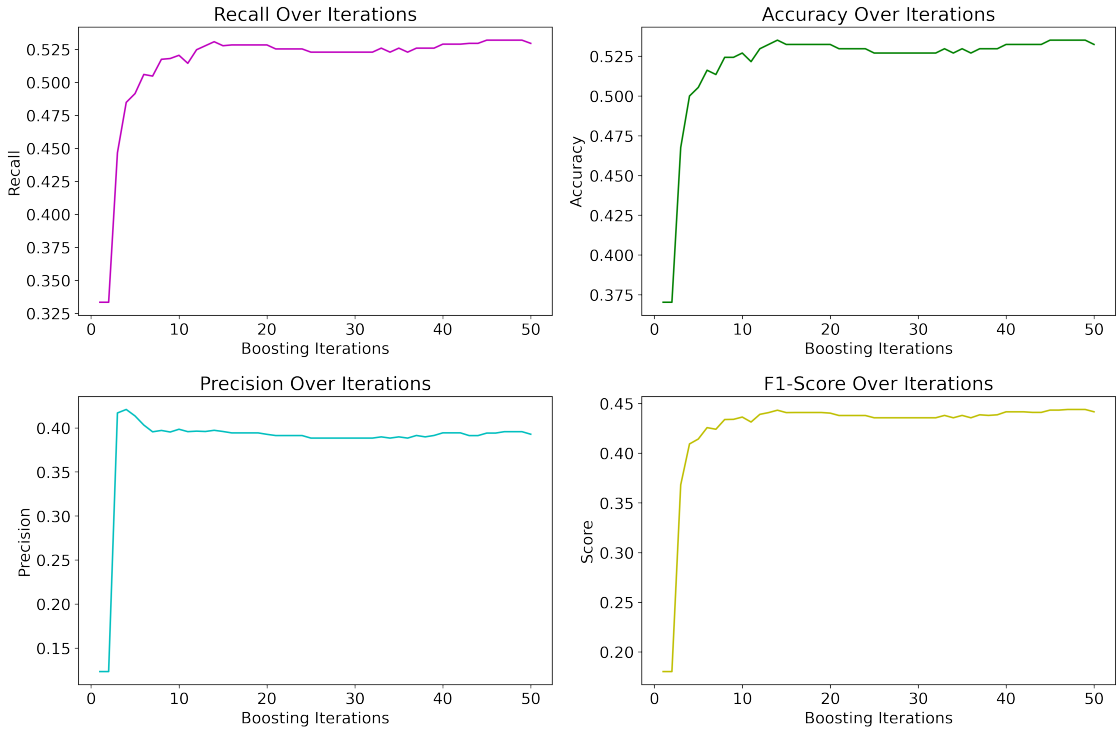


Figure 28. Performance Metrics Over Boosting Iterations.

Accuracy and Other Metrics:

- The model achieved an accuracy of 84.05% on the test data, which is a robust indicator of its general performance across all classes (Table[9]).
- Precision was measured at 84.14%, recall at 83.84%, and the F1 Score at 83.96%. These metrics attest to the model's balanced capacity for precision and recall, ensuring that it neither excessively misclassifies nor fails to detect relevant instances.
- The detailed breakdown by class showed that the model performed exceptionally well for the FISH_SWARM class, with a precision of 97% and a recall of 93%, highlighting its effectiveness in identifying more distinct swarm patterns.

Table 9. Performance Metrics of Each Class on Test-Dataset after Training.

	precision	**recall**	**f1-score**	**support**
FISH	0.81	0.82	0.82	137
FISH BEHAVIOUR	0.74	0.76	0.75	110
FISH SWARM	0.97	0.93	0.95	123
Accuracy			0.84	370
Macro Avg	0.84	0.84	0.84	370
Weighted Avg	0.84	0.84	0.84	370

5. Conclusion and future work

The study aimed at classifying underwater fish videos, focusing particularly on determining the presence of fish swarms. The primary objective of this study was to develop and validate an automated classification system capable of accurately identifying fish swarm behaviors from underwater video footage. This was successfully achieved through the implementation of a robust Fish Swarm model, underpinned by the Fish No-Fish (FNF) framework. The model demonstrated a significant capability in preprocessing video data, isolating frames containing fish and enabling efficient binary classification. The FNF model processed a substantial volume of video data, detecting fish in 2,108 out of 2,509 videos. This efficiency streamlined the data processing pipeline, focusing on pertinent video content and discarding irrelevant footage. This facilitated the development of a rich feature vector dataset, each vector encapsulating bounding box information essential for nuanced behavioral analysis, especially of swarming behaviors. The correlation analysis across the dataset revealed noticeable insights highlighting a strong correlation (0.86) between total detections and activity bursts, suggesting a link between frequent fish detections and heightened swarm activity, a notable positive correlation (0.56) between peak detection density and the class label, reinforcing the utility of this metric in differentiating swarm from non-swarm activities, and various other correlations such as between average and maximum frame coverage (0.74) indicate environmental or situational impacts on fish populations, while a negative correlation (-0.26) between minimum detection size and class label points to the challenges in correlating smaller detections with specific behaviors. Further insights from the FISH_SWARM dataset include the correlation (0.841) between total detections and activity bursts, emphasizing the relationship between higher fish activity and swarm dynamics, negative correlations involving minimum detection size, suggesting difficulties in accurately detecting smaller or less cohesive groups within swarms.

For the secondary objectives, advanced machine learning techniques, specifically Gradient Boosting Classifiers (GBC), were utilized to distinguish between swarm and non-swarm activities in underwater environments. The model's efficacy was demonstrated through feature extraction methodologies, leveraging spatial features derived from video frame analysis to ensure detailed and accurate input data for model training. The feature vectors encapsulated bounding box information essential for nuanced behavioral analysis, particularly swarming behaviors. To enhance the model's predictive capabilities, systematic hyperparameter tuning using GridSearchCV was employed. This process focused on optimizing parameters such as the number of estimators, learning rate, and tree depth. The

final model configuration, with a learning rate of 0.1, a max depth of 4, and 75 estimators, provided the best classification results.

The model underwent evaluation through methods such as k-fold cross-validation and train-test splits to optimize and verify performance across various metrics, including accuracy, precision, recall, and F1 scores. Before the validation phase, each class—FISH, FISH_-BEHAVIOUR, and FISH_SWARM—was represented with 617 instances, contributing to a total of 1851 data points, ensuring an unbiased assessment across all classes. This balanced approach is crucial for avoiding model bias towards more frequently represented classes. During the validation phase, the classifier exhibited robust performance, particularly in detecting fish swarm behaviors, achieving an overall F1-score of 87%. Notably, the F1-score for FISH_SWARM behaviors reached 97%, while for FISH and FISH_BEHAVIOUR classes, it was slightly lower at 82% and 80% respectively. These results underscore the model's proficiency in swarm detection but also highlight areas for improvement in classifying individual fish and fish behaviors.

These results underscore the need for further refinement of the model to improve its sensitivity and specificity across different classes of NON-SWARMS. The planned enhancement of the methodology for classifying non-swarm fish videos (fish and fish behavior classes) is designed to refine the detection and analysis of individual fish movements and behaviors. This new approach involves segmenting each video frame into three distinct parts: left, middle, and right. By structuring the analysis in this way, the model aims to more accurately capture and interpret the movements and positions of fish within their observed environment.

Suggestions for Future Work:

- Frame Segmentation: Each frame of the video will be divided into three sections. This segmentation allows for a focused analysis on different parts of the frame, potentially capturing variations in fish behavior that are specific to their position in the frame.
- Iterative Testing for Optimal Thresholds: To determine the most effective boundaries for each section, multiple iterations of testing with the dataset will be conducted. This process involves adjusting the boundaries and evaluating the model's performance to find the settings that yield the highest accuracy in behavior classification.
- Behavior Tracking Through Frame Analysis: By analyzing the behavior of fish from the first detected frame to the last, including key frames such as the first and last where fish are detected and multiple frames in the middle of the video, the model can more comprehensively understand the behavioral patterns. This approach leverages

temporal information that could indicate specific behaviors like entering or exiting the frame, sustained presence in a particular section, or quick transitory movements.

- Enhanced Feature Extraction: Utilizing the segmented frame approach, additional features related to the position, movement, and possibly even the orientation of the fish can be extracted. These features will be critical in differentiating between mere presence and significant behavioral actions such as aggressive interactions, feeding, or escaping.

The next phase of the project will be focusing on the spatial dynamics within segmented frames, implementing this enhanced methodology, conducting extensive tests to validate its effectiveness, and refining the algorithms based on the outcomes of these tests, and achieving higher precision in distinguishing between FISH and FISH_BEHAVIOUR classes. Continuous improvement through iterative testing and feedback will be crucial to developing a robust model that accurately reflects the complexities of fish behavior in various video contexts.

It might be interesting to see the recent developments in computer vision such as Attention Mechanisms [38] and Vision Transformers [39]. By incorporating ViTs into the classification pipeline for fish and fish behavior could offer a promising avenue to enhance model performance. Vision Transformers excel in managing spatial and temporal complexities within video data due to their robust attention mechanisms. This feature could refine the detection of subtle fish behaviors across segmented video frames, improving both accuracy and robustness against environmental conditions such as lighting changes and water turbidity. Additionally, extending ViTs to handle sequential video data, such as with adaptations like Video Vision Transformers [40] or TimeSformers [41], could capture dynamic behavioral changes over time more effectively. However, the implementation of ViTs will require assembling extensive datasets and may involve substantial computational resources. Future initiatives might focus on these aspects, leveraging advanced hardware or cloud computing solutions to facilitate efficient training and integration into existing systems. This approach will not only advance the technical monitoring capabilities of underwater environments but also deepen our understanding of fish behavior for better conservation and management practices.

6. Acknowledgements

I would like to express my deepest appreciation to my first supervisor, Prof. Jeffrey Andrew Tuhtan, for his professional supervision and unwavering support throughout every stage of my research journey. I am truly grateful for his confidence in me, from the moment he accepted me to embark on this research journey without hesitation, to his patience until the successful submission of my thesis without a shadow of doubt. I am also deeply indebted to my secondary supervisor, Elizaveta Dubrovinskaya, whose guidance and encouragement have been instrumental in navigating the challenges encountered along the thesis journey.

Particularly, I am thankful to my research members, Aleksandr Ivanov and Alexandra Kolosova, whose expertise and assistance have been invaluable in every discussion and suggestion. I would like to express my sincere thanks to them for their constructive feedback, support, and always being open to discussion, which has significantly enriched my research. I consider myself incredibly fortunate to have received mentoring from them during my master's studies, which have greatly contributed to my academic and personal development in the aquatic domain.

I am also grateful to my only love, Xin for her unwavering support and unconditional love, which have been the pillars sustaining my motivation and morale.

Lastly, I wish to give special thanks to my family and all my friends as a whole for their continuous trust and understanding throughout my master's studies and thesis research. Their steadfast support has been a consistent source of strength, grounding me through every challenge along the way.

References

[1] R. I. Perry et al. “Sensitivity of marine systems to climate and fishing: Concepts, issues and management responses”. In: *Journal of Marine Systems* 129 (2014), pp. 201–219. DOI: 10.1016/j.jmarsys.2013.09.009.

[2] Lisa G. Crozier and Jeffrey A. Hutchings. “Plastic and evolutionary responses to climate change in fish”. In: *Evolutionary Applications* 7.1 (2014), pp. 68–87. DOI: 10.1111/eva.12135.

[3] Olav Rune Godø et al. “Mesoscale eddies are oases for higher trophic marine life”. In: *PLoS One* 7.1 (2012), e30161. DOI: 10.1371/journal.pone.0030161.

[4] Roland Kays et al. “Terrestrial animal tracking as an eye on life and planet”. In: *Science* 348.6240 (2015), aaa2478. DOI: 10.1126/science.aaa2478.

[5] Amanda P. Sullivan, Douglas W. Bird, and George H. Perry. “Human behaviour as a long-term ecological driver of non-human evolution”. In: *Nature Ecology & Evolution* 3 (2019), pp. 168–174. DOI: 10.1038/s41559-018-0793-4.

[6] J. A. Fernandes et al. “Environmental effects of marine energy development around the world”. In: *Annex IV 2016 State of the Science Report: Environmental Effects of Marine Renewable Energy Development Around the World* (2017).

[7] N. J. C. Strachan. “Recognition of Fish Species by Colour and Shape”. In: *Image and Vision Computing* 11.1 (1993), pp. 2–10. DOI: 10.1016/0262-8856(93)90003-4.

[8] E. S. Harvey and M. R. Shortis. “A system for stereo-video measurement of sub-tidal organisms”. In: *Marine Technology Society Journal* 29.4 (1995), pp. 10–22.

[9] J. N. Fabic et al. “Fish Population Estimation and Species Classification from Underwater Video Sequences Using Blob Counting and Shape Analysis”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43.4 (2013), pp. 1006–1017. DOI: 10.1109/TSMC.2012.2226575.

[10] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. “Deep neural networks for object detection”. In: (2013).

[11] Zhong-Qiu Zhao et al. “Object detection with deep learning: a review”. In: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (2019), pp. 3212–3232. DOI: 10.1109/TNNLS.2019.2910418.

[12] Xiu Li et al. “Fast accurate fish detection and recognition of underwater images with fast r-cnn”. In: (2015), pp. 1–5.

[13] Xiu Li et al. “Accelerating fish detection and recognition by sharing cnns with objectness learning”. In: (2016), pp. 1–5.

[14] Xiu Li, Youhua Tang, and Tingwei Gao. “Deep but lightweight neural networks for fish detection”. In: (2017), pp. 1–5.

[15] James P. Horwath et al. “Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images”. In: *npj Computational Materials* 6.1 (2020), pp. 1–9. DOI: 10.1038/s41524-020-00387-5.

[16] Carl F. Sabottke and Bradley M. Spieler. “The effect of image resolution on deep learning in radiography”. In: *Radiology: Artificial Intelligence* 2.1 (2020), e190015. DOI: 10.1148/ryai.2020190015.

[17] George Cutter, Kevin Stierhoff, and Jiaming Zeng. “Automated detection of rockfish in unconstrained underwater videos using haar cascades and a new image dataset: labeled fishes in the wild”. In: (2015), pp. 57–62.

[18] Kaneswaran Anantharajah et al. “Local inter-session variability modelling for object classification”. In: (2014), pp. 309–316.

[19] Xavier Mouy et al. “FishCam: A low-cost open source autonomous camera for aquatic research”. In: *HardwareX* 8 (2020), e00110. ISSN: 2468-0672. DOI: https://doi.org/10.1016/j.ohx.2020.e00110. URL: https://www.sciencedirect.com/science/article/pii/S2468067220300195.

[20] Andreas Hermann, Jérôme Chladek, and Daniel Stepputtis. “iFO (infrared Fish Observation) – An open source low-cost infrared underwater video system”. In: *HardwareX* 8 (2020), e00149. ISSN: 2468-0672. DOI: https://doi.org/10.1016/j.ohx.2020.e00149. URL: https://www.sciencedirect.com/science/article/pii/S2468067220300584.

[21] Christian Haas et al. “Monitoring of Fish Migration in Fishways and Rivers—The Infrared Fish Counter “Riverwatcher” as a Suitable Tool for Long-Term Monitoring”. In: *Water* 16 (Jan. 2024), p. 477. DOI: 10.3390/w16030477.

[22] Kewei Cai et al. “A modified yolov3 model for fish detection based on mobilenetv1 as backbone”. In: *Aquacultural Engineering* 91 (2020), p. 102117.

[23] Chien-Yao Wang et al. “CSPNet: A new backbone that can enhance learning capability of CNN”. In: (2020), pp. 390–391.

[24] J.M. Hernández-Ontiveros et al. “Development and implementation of a fish counter by using an embedded system”. In: *Computers and Electronics in Agriculture* 145 (2018), pp. 53–62. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2017.12.023. URL: https://www.sciencedirect.com/science/article/pii/S0168169916311310.

[25] D. Zhang, X. Liu, and S. Chen. “Deep Learning for Fish Detection in Underwater Video Footage: A Comparative Study”. In: *IEEE Journal of Oceanic Engineering* 45.2 (2020), pp. 425–435. DOI: 10.1109/JOE.2019.2913664.

[26] Jürgen Soom et al. “Environmentally adaptive fish or no-fish classification for river video fish counters using high-performance desktop and embedded hardware”. In: *Ecological Informatics* (2022).

[27] H.-P. Fjeldstad, U. Pulg, and T. Forseth. “Safe two-way migration for salmonids and eel past hydropower structures in Europe: A review and recommendations for best-practice solutions”. In: *Marine and Freshwater Research* 69 (2018), pp. 1823–1837. DOI: 10.1071/MF18120.

[28] L. Hu et al. “Evaluation of Automated Image-Based Fish Detection in the Natural Environment: Implications for Research and Monitoring”. In: *Bioacoustics* 30.1 (2021), pp. 47–61. DOI: 10.1080/09524622.2020.1786739.

[29] Alfonso B. Labao and Prospero C. Naval. “Cascaded deep network systems with linked ensemble components for underwater fish detection in the wild”. In: *Ecological Informatics* 52 (2019), pp. 103–121. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2019.05.004. URL: https://www.sciencedirect.com/science/article/pii/S1574954118303078.

[30] Jun Hu et al. “Real-time nondestructive fish behavior detecting in mixed polyculture system using deep-learning and low-cost devices”. In: *Expert Systems with Applications* 178 (2021), p. 115051. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.115051. URL: https://www.sciencedirect.com/science/article/pii/S0957417421004929.

[31] He Wang et al. “Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and SiamRPN++”. In: *Computers and Electronics in Agriculture* 192 (2022), p. 106512. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2021.106512. URL: https://www.sciencedirect.com/science/article/pii/S0168169921005299.

[32] Usama Iqbal, Daoliang Li, and Muhammad Akhter. “Intelligent Diagnosis of Fish Behavior Using Deep Learning Method”. In: *Fishes* 7 (Aug. 2022), p. 201. DOI: 10.3390/fishes7040201.

[33] Jacek Maślankowski. “The evolution of the data warehouse systems in recent years”. In: *Zarządzanie i Finanse* 11.3, cz. 1 (2013).

[34] V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray Learning, 2013. ISBN: 978-1-84854-792-6.

[35] S. Kelling et al. “Taking a ’Big Data’ approach to data quality in a citizen science project”. In: *Ambio* 44.Suppl 4 (2015), pp. 601–611. DOI: 10.1007/s13280-015-0709-x.

[36] Glenn Jocher. *YOLOv5: An Improved Version of YOLO for Object Detection*. https://github.com/ultralytics/yolov5. Accessed: 2024-05-19. 2020.

[37] Jerome H. Friedman. “Greedy function approximation: A gradient boosting machine.” In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[38] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[39] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[40] Anurag Arnab et al. *ViViT: A Video Vision Transformer*. 2021. arXiv: 2103.15691 [cs.CV].

[41] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. *Is Space-Time Attention All You Need for Video Understanding?* 2021. arXiv: 2102.05095 [cs.CV].

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Afrasiyab Khalili

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "From Pixels to Patterns: Automated Classification of Fish Swarms in Underwater Videos", supervised by Jeffrey Andrew Tuhtan and Elizaveta Dubrovinskaya
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

20.05.2024

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.