Using Neural Networks to Automate Monitoring of Fish Stocks

Michael Stanley
Victoria University of Wellington
Wellington, New Zealand
stanlemich@ecs.vuw.ac.nz

Andrew Lensen
Victoria University of Wellington
Wellington, New Zealand
andrew.lensen@ecs.vuw.ac.nz

Mengjie Zhang
Victoria University of Wellington
Wellington, New Zealand
mengjie.zhang@ecs.vuw.ac.nz

***Abstract*—The fishing industry suffers from a lack of available data for governments to make informed decisions on how to best protect sea life and habitats. Manual monitoring is expensive, resulting in most caught fish going unmonitored on-vessel. The small sample size from the current approach is prone to information loss and biases. The application of machine learning in the fishing industry provides the opportunity to gain a broader understanding of the ecological impact of current fishing practices. This research explores the potential of deep learning in the context of computer vision to better monitor fish catch. It presents an approach to estimating the length of individual fish in millimetres from the output of a segmentation mask by a U-net-based neural network using an EfficientNet backbone. Our approach produced results that were, on average, within one centimetre of the measured lengths of fish, prior to any adjustment for halo effects of segmentation inferences.**

***Index Terms*—fish length estimation, real-world applications, image segmentation, EfficientNet, ResNet.**

I. Introduction

Globally, a lack of effective governance has seen the over-exploitation of many fish stocks [1], bringing them ever closer to the risk of collapse. Research from the Minderoo foundation (2021) [1] on the sustainability of marine fisheries in 142 coastal states found that 49% of assessed fish stocks are now over-fished, with almost 10% having been driven to collapse. Research conducted by the Food and Agriculture Organisation (FAO) in 2020 [2] estimated global fish production to have reached about 179 million tonnes in 2018, with total fish consumption rising 122% from 1990, and fish consumption is estimated to have provided 3.3 billion people with almost 20% of their average per capita intake of animal protein [2]. Meanwhile, fish stocks within biologically sustainable levels dropped from 90% in 1990 to 65.8% in 2017. The situation is likely to be even worse for unassessed fisheries [3].

Science-based management of commercial fisheries suffers from a lack of availability of data [1]. For informed decisions to be made, governments need greater insight into the state of sea life. Neural Networks applied in computer vision offer the opportunity to gain information on current practices without introducing biases or information loss at the catch stage [4].

Advances in neural networks and the increase in speed of the GPUs required to run them over the last decade have allowed the practical implementation of this technology in a wide range of industries. However, the adoption of smart monitoring systems in the fishing industry in New Zealand is yet to be implemented. Availability of data surrounding the fishing industry and their onboard activities is not widely available; visual data of the deck is often locked behind privacy concerns where employees of the vessels may be identified, and all imagery is reliant on having the approval of either the captain or the organisation to be distributed. As such, the analysis of the potential of these techniques in the industry has been hampered.

Non-invasive approaches to monitoring are crucial when applying automated systems on-vessel; the privacy of those working on-vessel is highly important. In New Zealand, data collection for monitoring the types and quantity of fish caught is mainly reliant on human reviewers [5]. With over four million square kilometres of ocean, efforts to monitor the multi-billion dollar New Zealand fishing industry are difficult and expensive, resulting in only around 25% of deepwater catch occurring on a vessel with an observer [5].

In this paper, we explore the use of automated image processing techniques such as deep neural networks (NNs) and investigate their potential for monitoring fish on-vessel. Data consists of imagery of tarakihi, a species of fish that has been identified to be below its soft limit in New Zealand waters. As part of this project, data was gathered manually to be explicitly used in tackling the problem of automating the monitoring and length estimation of the tarakihi fish.

This paper aims to investigate the application of deep learning for performing fish (Tarakihi) length estimation with a very small amount of input data. The use of supervised networks for data sorting as an approach for reducing the level of manual data cleaning required is also explored. Specifically, we will:

1) Discuss the data collection approach, including the careful considerations required to ensure capturing a significantly diverse and representative dataset;
2) Investigate the application of EfficientNet [6] for automating the process of locating fish in the imagery dataset and the use of interpretable techniques for evaluating the inferred contours.
3) Propose an approach for inferring the length of the detected fish. This approach will be evaluated across the collected data to establish the accuracy and robustness of the proposed approach.

II. Background

Declining fish stocks and damage caused to sea habitats from large-scale fishing practices have led to the development of strict regulations regarding where fishing may be conducted, the quantity of fish that may be caught, and the size of fish that may be kept. Size limits have been imposed to help protect fish stocks, allowing fish to grow such that they are given a chance to breed at least once [7].

Prior research into the use of machine learning-based computer vision techniques in the fishing industry has seen some success. Monkman et al. [8] investigated object detection on side-face images of European sea bass and were able to predict the location of fish in images with a mean intersection over union (IOU) of 93%. In this research, fiducial markers of varying sizes were used for calibrating images and provided a point of reference when measuring the length of the fish. Raw imagery in this dataset consisted of the side profile of fish, imagery was horizontally aligned with the fish. This allowed for simple calculation of fish length based on the bounding box and the use of a calibration object in each image. Such an approach provides insight into how length estimating may be tackled. However, this method will not be directly translatable to many real-world use cases. Imagery datasets rarely consist of such favourable examples where a single axis may be used to estimate length.

Work by Qiu et al. [9] investigated the use of transfer learning for image classification on a popular fish dataset, Croatian and QUT, in 2018. Their use of pre-trained bilinear convolutional neural networks showed some success by improving the accuracy of popular networks on datasets that are particularly challenging due to the poor quality of the imagery. However, the computational load was significantly increased. As classification models output only class-specific data, geometric properties for length estimation are not present and would therefore provide little benefit at the current stage of this research.

Stereo optical systems are relatively popular in the use of monitoring fish in water; the National Oceanic and Atmospheric Administration have used baited remote underwater video stations (BRUVS) equipped with a stereo video system for at least ten years [10]. The use of stereo video systems opens up the opportunity for distance to be measured. Rodriguez et al. [11] paired this technology with the technique of background subtraction to segment the fish imagery into foreground and background. The output of this process was a binary representation of the candidate foreground from which size, area, and length-to-height ratios are used to determine whether the object is a fish. In practice, stereo camera systems are not present on commercial fishing vessels due to their added cost and maintenance requirements.

Álvarez-Ellacuría et al. [11] sought to use a different approach for the estimation of the length of the European hake. Polygon-based image segmentation was used to identify individual fish in images. This provides an advantage over bounding boxes as polygons are not bound to just two axes, allowing for rotation-invariant estimation of length. Their data consisted of images of overlapping hake in boxes, meaning many of the fish were not fully visible. To combat this, they segmented the heads of fish and used known head-to-total length relationships to extrapolate the total length per fish detected. This approach was able to identify 87% of heads in the imagery. However, their approach was sufficiently accurate for real-world deployment.

More recently, Palmer et al. [12] also investigated the use of image segmentation to measure the number of obscured fish in buckets. Their study focused on the common dolphinfish, an important part of the commercial fishing industry in the Mediterranean. Their use of high-resolution imagery and a larger dataset was able to achieve an accuracy of 86.10% on a dataset consisting of 4117 fish from 276 images. As the weight of each bucket was also measured, researchers were able to use the bucket weight and the predicted number of fish in each bucket to estimate the mean fish length per bucket. Deviations between observations and estimates ranged between -7.4 and 4.8 centimetres. The weight of fish is not practically available on fishing vessels, as fish must be processed in real-time, with bycatch immediately discarded overboard.

Deployment of real-time neural network-based approaches on fishing vessels (edge computing) will require models with low computational cost. EfficientNet-based models show considerable promise, given their very competitive performance on problems such as ImageNet despite their significantly smaller size [6].

Where this research differed from past literature is that fish were removed from the buckets they are auctioned from, meaning that where other research had to approximate length from the head size [13] or average length per box [12], this research focused on directly measuring the length of individual fish from imagery. Interpretable techniques are also explored for their potential in explaining which shape features of contours found from segmentation masks may provide insight into identifying contours of high or low quality, reducing the likelihood of poor inferences affecting recorded lengths. The use of interpretable techniques [14] provides justification for the inferences made by ML models, the features that contributed to individual predictions may be visualised in the form of a contribution score. By no longer being a "black box", a greater trust may be placed in these systems and their reasoning understood.

III. Overall Method

The first stage of this research was to gather imagery data. This data had to be gathered in a manner that may be replicated onboard fishing vessels in New Zealand, it also had to accommodate for a variety of camera angles and distances, as the specific set-up amongst the vessels is inconsistent. Data from on-vessel cameras is highly variable, often noisy or of poor quality, as cameras are subject to harsh conditions while out at sea. As such, an approach that is invariant to scale and rotation must be considered. Artificial intelligence provides this opportunity and its potential for predicting the size of

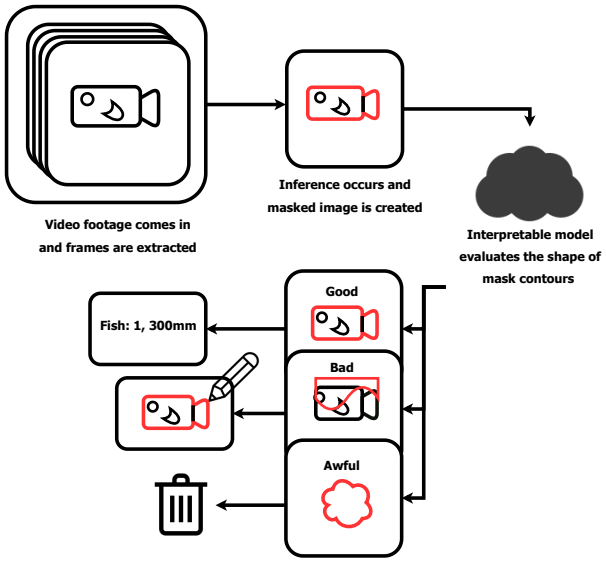


Fig. 1: Data Cleaning process visualisation.

the tarakihi fish, an overfished species in New Zealand, was explored in this research.

To facilitate the efficient collection and processing of data, the use of machine learning for sorting inferences based on the shape of inferred objects was also investigated. Figure 1 illustrates this process, showing that images are first inferred on by the segmentation model and contours identified of inferred objects are then subject to an evaluation using their shape features.

A subset of the videos was chosen to be annotated for segmentation, these included videos from both a fixed and handheld camera allowing for a varied train set. Polygons were then drawn on the frames extracted from this data that contained fish. Only images in which the fish had been moved or a new fish had entered the frame were annotated. This helped to increase the information gain of each annotated image by avoiding annotating consecutive frames which would be visually similar.

Shape features were calculated from both quality fish polygons and the polygons of false positive inferences to train an Explainable Boosting Machine (EBM). The shape functions of objects inferred from subsequent runs of the inference model were then evaluated by the EBM. The Contours with a circle deviation within the range of values associated with fish, that were not predicted to be fish by the EBM, were targeted for further training.

A holdout set of images were then inferred on by the segmentation model. Segmented objects that were classified as having quality contours also had their pixel length calculated. This was done by drawing a minimum bounding circle to enclose the contour of the object, The diameter of this circle was used as the object's pixel length. By locating a checkerboard pattern in the image and comparing its pixel length to the known length, the scale factor for the current image was calculated and the predicted millimetre length of segmentation inferences was calculated.

Each of these stages is discussed in detail in the following sections.

IV. Data Collection & Processing

The data collection conducted in this research was an iterative process. The different approaches taken were not done just to collect the information to be used in training computer vision learning algorithms but also to provide evidence on the methods required to most accurately infer the length of a fish from a single camera. Multiple factory visits were conducted and two distinct methods for collecting the data were used.

In the first visit, data were collected using a handheld camera facing top-down over fish as they were passed over a metal table. While in practice, the camera would be fixed on the vessel, by holding the camera we were able to simulate the vessel's movement and gather data from various angles and distances. The goal of this approach was to reduce overfitting to a particular setup.

Though the approach of using a free camera was able to provide us with a dataset that contained a large variety of different observation angles for the measured fish, there were also significant drawbacks. One such drawback was an inconsistent and highly variable scale factor in the images derived from this approach. The scale factor was a measured value for the known difference between pixel and millimetre distance, found by measuring the checkerboard calibration pattern. Depending on the position of the pattern relative to the fish the derived scale factor would lead to an over or under prediction of the true scale factor at the position of the fish. Using a free camera also meant that the scale factor had to be recalculated for every image, as the camera angle and distance were always moving.

The second method for gathering data utilized a fixed camera. A metal frame was used to hold the camera stationary above the table, allowing for a more consistent scale factor. This fixed perspective also enabled us to transform the images ensuring that only the region of interest was visible.

The camera used in this research was a custom-built camera used specifically for fishing vessels. It had a resolution of 960x540 and recorded imagery at 15 frames per second. Data collection was conducted in factories where access to many samples was possible. A total of 21 buckets of tarakihi were used throughout this research. A checkerboard pattern on non-reflective, paper was placed on the metal table for camera calibration and as a point of reference for length estimation.

One challenge was simulating fish in motion. Live samples of fish were unavailable and the samples needed to be placed in positions representative of imagery that may occur with fish moving down a chute or being passed at high speeds during sorting on a vessel. As shown in Figure 2, fish in motion present a significant amount of motion blur, making a precise length estimation more difficult. The fish used in this research had been frozen whilst stacked in the boxes, which helped to simulate the curling shape of live fish. In total 42 videos

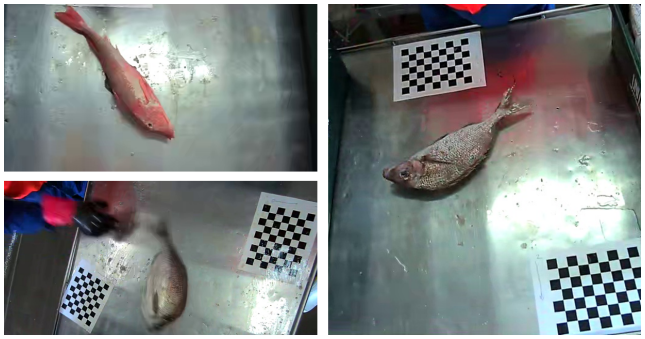


Fig. 2: Examples of imagery taken. The top left is a gurnard. The bottom left is a tarakihi in motion. The right is a clear example of tarakihi.

were gathered, these videos ranged in lengths of up to 15 minutes. Around 170 thousand frames, in total, were extracted from these videos to create the dataset for this research, though many did not contain images of tarakihi. This data had to be sorted and data that was not usable was removed. To reduce projection errors that may affect the perceived size of objects when converting a three-dimensional scene to a two-dimensional image plane; a bird's eye transformation was conducted on images taken from the fixed camera. This approach brought other benefits also, such as reducing the number of fish present in a scene down to a single individual. The edge points of the table on which fish were measured were chosen as our control points for the limits of our warped region. The width and height of the output image were calculated by using Pythagoras' Theorem to identify the maximum hypotenuse, in the image coordinate system, for both the width at the bottom and top points and the height on the left and right of our projected table. The OpenCV function CV2.GetPerspectiveTransform [15] uses both the edge points from the original image and the max width and height from our desired region of interest to calculate a 3x3 perspective transformation matrix [15]. This matrix was then applied to the image providing us with our transformed image containing only the region of interest.

V. Fish Segmentation

The VGG annotation tool [16] was used to annotate the training data for segmentation. The tool was developed by researchers at Oxford and provides a simple and intuitive open-source environment, hosted through a browser, for drawing a variety of annotations such as bounding boxes and polygons.

Polygons were drawn around both the fish in the centre of the frame and fish in buckets if visible. The training was targeted to maximise the information gained from each annotated sample. The images were extracted from 15 frames per second camera feed. Therefore, there were many frames that were visually identical, and annotating many that were close together would yield little benefit when trying to explain the complexity of all the data. Instead, images were only annotated if a fish had been manually changed, this may include a flip or rotation and transformation.

2020 images were annotated with some of these images containing multiple fish polygons. Images were split randomly into a train and test set at an 80%/20% ratio. A holdout set was created manually prior to this split and contained 301 images, with 41 images from the first factory visit, 116 from the second, and 144 images from the third visit. This holdout set was used only to evaluate the performance of segmentation inferences while a second holdout set of non-annotated imagery was used for evaluating the performance of the proposed method for estimating length. From the data gathered in the third factory visit, only images of fish from the first bucket were annotated, and fish from the remaining five buckets were used in the holdout set for evaluating estimated lengths.

Training images were augmented by applying horizontal flips, rotation, brightness, and gamma transformations as well as the addition of Gaussian noise to reduce overfitting. The output of the segmentation model was a two-dimensional array with values between zero and one for the predicted confidence that a pixel, at the corresponding coordinates in the input image, contained a fish, this was multiplied by 255 when saving the array as a mask image.

The segmentation model was a Unet-based model with an EfficientNet-B3 backbone. The canonical EfficientNet paper [6] used neural architecture search to develop a feature network topology that allowed for significantly improved performance with a comparatively small model, by using a bi-directional feature pyramid network (BiFPN). Given the model's efficiency, we kept images at (320×320 pixels).

This model was trained for a single class for 60 epochs with a batch size of eight and a learning rate of 0.0001. EfficientNet-B3 was chosen as it is the highest performing EfficientNet model before there is a significant trade-off between performance and the number of parameters [6]. Although the model size and inference speed are not of concern in this research, emphasis was given to smaller systems that may be capable of running on edge devices out at sea or capable of providing an observer with close to real-time assessments of the catch.

Close to real-time assessment of catch may provide valuable statistics for making more informed decisions about where to fish and result in less bycatch and less damage being caused to aquatic creatures and their habitats. Such methods should not only be limited to the assessment of sea-life brought onto the vessel but also to sub-marine systems in the vicinity of fishing vessels, such as the increasingly common trawl camera, attached to deployed trawl nets [17] which provides a live feed of the fish as they are caught during a trawl.

VI. Interpretable methods for evaluating contours

The contours of inferred objects were used to evaluate the quality of inferences and reduce the likelihood of poor inferences affecting estimated lengths. By using shape features derived from these contours a generative additive model (GAM) was trained to classify high and low-quality inferences.

Contours that were classified as low quality were used to flag images that required further training or potential false positives. While those of high quality were used in the process of estimating length.

OpenCV's findContours function [18] was used to identify these and describes contours simply as a curve joining all continuous points along a boundary with the same colour or intensity.

The target contour in the image was found by comparing the areas and distance from the centre of each contour. From the contours with an area, in pixels, between 4,000 and 60,000 the closest to the centre was selected as the contour most likely to be the fish currently being captured. This contour was then further analysed to predict its quality.

A separate mask was generated containing only the inference in the centre of the image. Various features were calculated from the contour found from the remaining inference and were analysed using interpretable machine learning techniques developed in research conducted by Nori et al. [14].

Two datasets were created for evaluating this system, one for high-quality segmentation masks, and another for low-quality segmentation masks. Segmentation masks were deemed high-quality if they were of a standard suitable for training, this meant that the entire fish would need to be included in the mask and minimal false positive pixels would need to be present.

Low-quality segmentation masks were a broader category. Masks unsuitable for training were included, as well as masks from false positive inferences or masks generated from multiple fish overlapping that did not represent the shape of an individual fish. This allowed us to reduce the chances of using an inferred segmentation of multiple fish for a length estimation.

The algorithm was then trained on various features calculated from these two datasets. Some features were raw inputs derived from OpenCV's contours [18], others were calculated by using geometric statistics common in particle analysis, and remote sensing. Input features were gradually reduced until there were only three. Fewer inputs meant that inferences were always justified based on a few input features and were more easily interpreted and less likely to be over-fit to the training data.

InterpretML [14] was chosen as it provides insight into both global and local features. Research into interpretable machine learning has been gaining traction in recent years [19]. Interpretable methods allow for greater trust in the systems in which they are integrated, as decisions are less likely to be challenged if their justification is clear. By using interpretable machine learning techniques, not only may a robust automated system, that is capable of determining the quality of polygons, be developed, but the contributing factors that resulted in this prediction may be easily understood and visualised.

As part of InterpretML research [14], a new algorithm was developed with a focus on interpretability, the Explainable Boosting Machine (EBM). The EBM is a generalised additive model (GAM) and uses modern machine learning techniques to learn feature contribution functions. This function determines the contribution to either the positive or negative class at each value of the input feature.

As EBM is an additive model feature scores are easily interpreted due to their modular contribution to the prediction [14]. The contribution of each feature may be visualised by plotting the feature contribution scores identified when training the EBM, an example is shown in Figure 3. This non-linear relationship is called the shape function in the research conducted by Lou et al. [20].

Global scale interpretability provides the importance of each feature, this is calculated by using the mean absolute score, allowing us to ascertain which features are most important in determining whether a contour is highly likely to be that of a fish or not.

On the local scale each input feature has a score to determine its impact on the predicted class, this score can be both positive; increasing the likelihood that the observed data belongs to the positive class, or negative; decreasing this likelihood.

Several features were calculated from the contours found in the segmented image, in the form of shape factors, to better evaluate their shape. Shape factors are used for analysis in a variety of fields such as image analysis and microscopy, where the shapes of objects must be differentiated. The notable geometric properties of fish contours that were explored included circularity, compactness, and complexity. These were the input features identified as having the highest feature importance when conducting iterative feature reduction when training the EBM. 35 Features were first included, and those with low feature importance or high collinearity were gradually dropped until three remained; circle deviation ($circ\_dev$), elongation ($elong$), and coordinate complexity ($complex$). Podczeck (1996) [21] explored the use of shape factors to assess the shapes of particles. As part of this assessment, three of the considered parameters for describing shapes analysed the deviations from standard shapes such as a square, triangle, and circle. The deviation from a circular image ($circ_dev$) was included in our analysis of viable contours as an early exploration of contour features showed it to be highly important in determining quality fish contours.

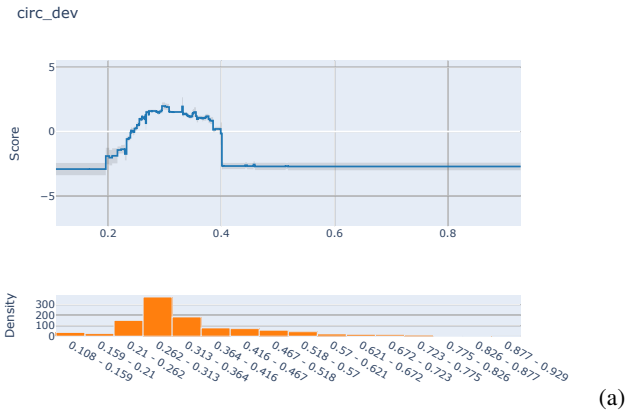
$$circ\_dev = \frac{area}{\frac{\pi}{4} s_m^2} \tag{1}$$

Where s_m is the longest side of a minimum enclosing rectangle.

Two additional features were included in the final explainable model and are calculated using the characteristics of the external border of our shape and the pixel length of the contour. These were: Podczeck's elongation calculation [21]

$$elong = \frac{perimeter}{length} \tag{2}$$

The second equation was an adaption of Podczeck's elongation, which instead explains the number of vertices relative



(a)

Fig. 3: Shape function plotted for circle deviation

to the length of the shape, the name coordinate complexity is used to refer to this feature and it was calculated as.

$$complex = \frac{ext.coords}{length} \tag{3}$$

The data used in training the explainable boosting machine had to be carefully selected using a mixture of both inferred and manually created segmentation masks. 160 of the masks included in the positive class were manually drawn, and the remaining 1205 masks were inferred. In earlier iterations, human-made masked images made up roughly 50% of the positive class. A larger number of masked images that were created from human annotated polygons were included in the training data to ensure quality contours were being used. However, this resulted in inflated feature importance being given to the number of exterior coordinates, as human-drawn polygons generally have far fewer exterior coordinates than polygons from an inference.

Circle deviation ranges between zero and one, values close to one signify our contour is highly similar to a circle whereas values close to zero signify a very dissimilar shape. The circle deviation graph 3 shows the contribution scores found from the training set of contour features. The range of values for the circle deviation is shown on the x-axis, while the contribution is shown on the y-axis. Illustrated below the graph is a density plot showing the number of instances for each range of feature values. Contours with a circle deviation value of less than 0.24 and greater than 0.4 had a similar score of -2.41, suggesting these shapes are not likely to be fish. Circular deviation scores between 0.24 and 0.4 had a positive contribution to the prediction that our shape is that of a quality fish segmentation.

A second check was created, based on findings from the shape functions, for contours that did not pass the EBM. A contour that failed the first check but still had a $0.25 < circdev < 0.39$ was considered to potentially have a clearly visible tarakihi. This was done as a proxy for entropy-based sampling to target training at these examples where contours were similar but did not pass the EBM's inference. This range of circle deviation values was chosen as there was a large number of observations in this range, so the inference that contours within this range are likely to be those of a fish is based on many examples. Circle deviation was also measured to have the highest feature importance for identifying quality fish.

The first run of this system placed 425 images into the 'bad' output directory. 19 of the images classified as containing bad fish masks were incorrect (false negatives). 965 images failed the EBM evaluation but had a circle deviation value within the range that was considered for further evaluation, these were placed in a directory for later annotation. 1601 images were placed in the good segmentation category, 257 of these were considered to be false positives as the segmentation did not cover the entire fish.

VII. Length Estimation

The length of fish was derived from the inference mask. A binary thresholding approach was used to remove segmentation artefacts by setting all values above the threshold, of 128, to 255 and all other values to 0. OpenCV was then used to create a minimum bounding circle around the fish, the diameter of the circle was used as the pixel length.

The checkerboard pattern was used to derive the pixel-to-millimetre ratio. An adaptive approach was used to search for partially-visible checkerboards. The pixel lengths of all four sides of the checkerboard were then calculated based on the size of the grid that was found in the image. Estimates for the number of millimetres per pixel were calculated by getting the average of each side's length divided by its pixel length. This value was multiplied by the pixel length of the fish to give its length in millimetres. The pixel to millimetre ratio for a single side of the checkerboard pattern was calculated with the following equation:

$$R = \frac{L_{mm}}{\sum_i N_{pix}} \tag{4}$$

where the length in millimetres is denoted by L_{mm} and $\sum_i N_{pix}$ is the sum of the pixel distance for the given side (i).

The length calculation for an individual fish from the diameter of the minimum enclosing circle in pixels D_{pix} using the pixel to millimetre ratio from each of the four sides $\sum_i R$ of the checkerboard pattern is calculated as follows:

$$L = D_{pix} * \frac{\sum_i R}{4} \tag{5}$$

VIII. Deriving Lengths from Videos

In order to effectively measure the length of an individual over multiple frames, lengths inferred from the automated process would need to match up with those lengths measured on-site, when ordered chronologically. This would mean that double counting or missing fish would have a significant adverse effect when comparing true and predicted lengths. Partial inferences when hands were covering fish were one instance that would significantly alter measured lengths.

After inference was conducted on all images from the holdout set, masks were ordered chronologically; by using the UNIX timestamp of the video from which the image was extracted and the images corresponding frame number. This ensured that important temporal information that may assist in the evaluation of predicted lengths was not lost when reading frames and masks. The median of similar lengths was used to determine the estimated length of the current fish, the median was chosen as false positives could skew the recorded max length if the measured contour significantly increased in size. Lengths were considered similar if they were within 1cm of the median length. If an inferred length was greater than one centimetre above that of the median estimated length for the current fish, this value was then stored, and only if there are six consecutive frames with a similar length is this considered to be the new estimated length for the current fish. Allowing lengths to always increase in size but not decrease enabled us to reduce length measurements from partially visible fish affecting the recorded length of an individual. The estimated length was set to zero after consecutive frames where no fish was present to ensure that the estimated lengths of one individual would not affect the recorded length of the following fish.

IX. Results

Fish used in the holdout set for evaluating the length estimation process were measured on-site with mm accuracy, this dataset was made up of 129 individuals. The dataset was reduced to just over 55 thousand image frames that were extracted from the raw footage gathered on-site. This was done only to include the images in which fish were present, removing the need to infer on many frames where no fish were in view. The segmentation model inferred on these images and prediction masks were cropped below the ruler on which fish were measured. As all fish were measured before being placed on the table, cropping inferences in the region where fish were measured allowed us to simulate a pause between observed fish. This pause, of around 60 frames, was used to distinguish when a new fish was entering the scene. Lengths from the holdout set from the third visit were then measured using the aforementioned process and a CSV file was outputted with a single predicted length (the median) for each observed individual. The Lengths from this CSV were then compared to the ground truth recorded lengths, these values were adjusted for unknown error, and are illustrated in Figure 4.

To correct for unknown error that may affect the predicted lengths a linear regression for the predicted lengths from the second visit was conducted providing us with a linear transformation for future predicted lengths to be adjusted. A single adjusted length was calculated as:

$$L_i = P_i * 0.9196 + 45.593 \tag{6}$$

Equation 6: Linear adjustment for predicted values. Where L is the adjusted length for fish i and P is the predicted length for an individual derived from a video. The adjusted lengths are recorded for each fish and compared to the true lengths in Figure 4.

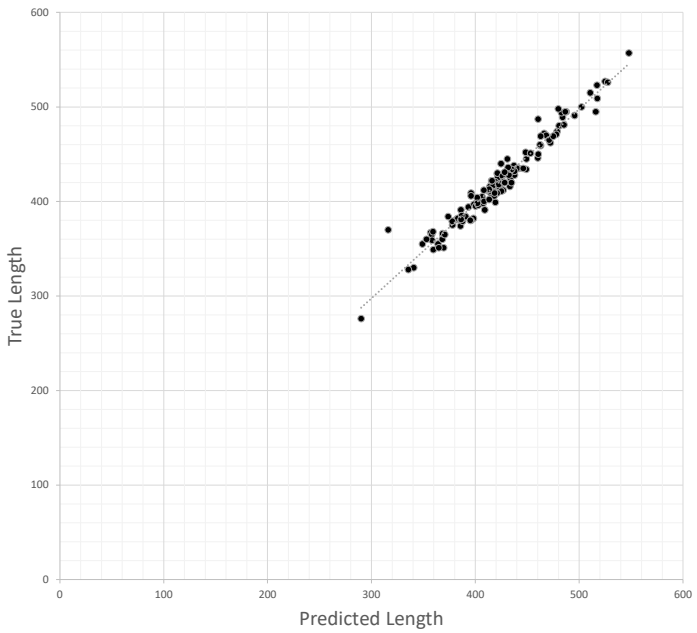


Fig. 4: True lengths compared to adjusted predicted lengths.

Figure 4 shows that this method for estimating lengths produces similar values as those measured manually. The average absolute difference between adjusted predicted and true lengths, for recorded fish from the third visit, after this adjustment was 7.523 millimeters. The R-Squared of the linear relationship was 0.9536, suggesting the predicted lengths closely fit those measured on-site. The difference between inferred lengths ranged from as low as 0.12 millimetres above the recorded length to 75.85 millimetres above the recorded length.

X. Conclusions and Future Work

This paper investigated the potential for an automated system to measure the length of tarakihi fish from images by using machine learning and computer vision techniques. Two distinct methods for collecting the data were presented and compared. The use of a handheld camera was able to provide us with a more varied dataset, which was beneficial for training. However, a fixed camera allowed for a more consistent scale factor and was better suited for producing accurate length estimates. An EfficientNet model was trained to produce segmentation inferences and mask images were created from these to identify pixels containing fish. By using a checkerboard pattern as a point of reference pixel lengths, from a minimum enclosing circle of the inferred object, were converted to millimetres. The shapes of these inferences were evaluated using an EBM to predict their quality based on shape features. This allowed fewer poor to be considered when calculating the median predicted length.

Results show that fish length estimation from segmentation inferences is possible, and was on average within one centimetres of the measured lengths. However, it is clear further research is required before such a system may be implemented. The use of an enclosing circle was used to determine the length of an inferred fish successfully, but a larger dataset of measured lengths gathered on vessels is needed to evaluate the model's performance in environments beyond the controlled ones created for these tests. Future work will also explore the use of instance segmentation for counting individuals and storing information about recently seen fish that may have left and re-entered the scene. The current approach for counting and measuring has no resilience against fish that are not manually singulated. Fish that are close to one another in an image would share the same contour and result in a single length estimation which may be unrepresentative of the length of either individual.

Currently, the standard in the industry is to measure the length of tarakihi by their fork length. Though not covered in this research, identifying where the tail forks may be a challenge for future work alternatively the accuracy of extrapolating fork lengths from full length may be explored. Changes in length estimation based on the distance and angle of the calibration pattern from the camera is another area that may be beneficial to explore. The use of a per-tile pixel to millimetre ratio would provide insight into how objects in different areas of the image scale relative to the pattern and may provide us with more accurate measurements.

XI. Acknowledgements

This work was a collaborative effort between Victoria University, Wellington, and Lynker Analytics. I would like to thank David Knox, Lynker Analytics, for his mentorship and encouragement throughout.

References

[1] Minderoo Foundation. "New report finds global fisheries in far worse state than previously estimated." (2021), [Online]. Available: https://www.minderoo.org/global-fishing-index/news/new-report-finds-global-fisheries-in-far-worse-state-than-previously-estimated/.

[2] Food and A. Organization, "The state of world fisheries and aquaculture (sofia)," *Science*, p. 244, 2020.

[3] C. Costello, D. Ovando, R. Hilborn, S. D. Gaines, O. Deschenes, and S. E. Lester, "Status and solutions for the world’s unassessed fisheries," *Science*, vol. 338, no. 6106, pp. 517–520, 2012.

[4] Y. Lin, T. Hung, and L. T. Huang, "Engineering equity: How AI can help reduce the harm of implicit bias," *Philosophy and Technology*, vol. 34, no. 1, pp. 65–90, 2020.

[5] Ministry for Primary Industries. (), [Online]. Available: https://www.mpi.govt.nz/fishing-aquaculture/sustainable-fisheries/strengthening-fisheries-management/monitoring-observing-fishing-activity/.

[6] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. arXiv: 1905.11946.

[7] M. for Primary Industries. (), [Online]. Available: https://fs.fish.govt.nz/Page.aspx?pk=130&tk=580.

[8] G. G. Monkman, K. Hyder, M. J. Kaiser, and F. P. Vidal, "Using machine vision to estimate fish length from images using regional convolutional neural networks," *Methods in Ecology and Evolution*, vol. 10, no. 12, pp. 2045–2056, 2019.

[9] C. Qiu, J. Cui, S. Zhang, *et al.*, "Transfer learning for small-scale fish image classification," in *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, 2018, pp. 1–5.

[10] NOAA. "Baited remote underwater video station (bruvs) surveys of fish in the hawaiian archipelago from 2012 to 2014." (2014), [Online]. Available: https://www.fisheries.noaa.gov/inport/item/25248.

[11] A. Rodriguez, A. J. Rico-Diaz, J. R. Rabuñal, J. Puertas, and L. Pena, "Fish monitoring and sizing using computer vision," in *Bioinspired Computation in Artificial Systems*, J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, F. J. Toledo-Moreo, and H. Adeli, Eds., Cham: Springer International Publishing, 2015, pp. 419–428.

[12] M. Palmer, A. Álvarez-Ellacuría, V. Moltó, and I. A. Catalán, "Automatic, operational, high-resolution monitoring of fish length and catch numbers from landings using deep learning," *Fisheries Research*, vol. 246, p. 106 166, 2022.

[13] A. Álvarez-Ellacuría, M. Palmer, I. A. Catalán, and J.-L. Lisani, "Image-based, unsupervised estimation of fish size from commercial landings using deep learning," *ICES Journal of Marine Science*, vol. 77, no. 4, pp. 1330–1339, Nov. 2019.

[14] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretml: A unified framework for machine learning interpretability," *arXiv preprint arXiv:1909.09223*, 2019.

[15] *Geometric image transformations*.

[16] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19, Nice, France: ACM, 2019.

[17] N. Bazilchuk, *An in-trawl camera for fish*, Apr. 2015.

[18] OpenCV. (), [Online]. Available: https://docs.opencv.org/3.4/d4/d73/tutorial_py_contours_begin.html.

[19] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1900654116.

[20] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12, Beijing, China: Association for Computing Machinery, 2012, pp. 150–158.

[21] F. Podczeck, "A shape factor to assess the shape of particles using image analysis," *Powder Technology*, vol. 93, no. 1, pp. 47–53, 1997.