Automating the assessment of biofouling in images using expert agreement as a gold standard

Nathaniel J. Bloomfield[1,*]**, Susan Wei**[2]**, Bartholomew Woodham**[3]**, Peter Wilkinson**[3]**, and Andrew Robinson**[1]

[1]Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne
[2]Mathematics And Statistics, The University of Melbourne
[3]Biosecurity Animal Division, Department of Agriculture, Water and the Environment

ABSTRACT

Biofouling is the accumulation of organisms on surfaces immersed in water. It is of particular concern to the international shipping industry because fouling increases the drag on vessels as they move through the water, resulting in higher fuel costs, and presents a biosecurity risk by providing a pathway for marine non-indigenous species (NIS) to establish in new areas. There is growing interest within jurisdictions to strengthen biofouling risk-management regulations, but it is expensive to conduct in-water inspections and assess the collected data to determine the biofouling state of vessel hulls. Machine learning is well suited to tackle the latter challenge, and here we apply so-called deep learning to automate the classification of images from in-water inspections for the presence and severity of biofouling. We combined images collected from in-water surveys conducted by the Australian Department of Agriculture, Water and the Environment, the New Zealand Ministry for Primary Industries and the California State Lands Commission, and annotated them using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. We compared the annotations from three biofouling experts on a 120-sample subset of these images, and found that for two tasks, identifying images containing fouling, and identifying images containing heavy fouling, they showed 89% agreement (95% CI: 87–92%). It was found that the MTurk labelling approach achieved similar agreement with experts, which we defined as performing at most 5% worse than experts (p=0.004–0.020). If used in future this approach could offer significant cost savings over relying on expert consultants to annotate similar data. Our deep learning model trained with the MTurk annotations also showed reasonable performance in comparison to expert agreement, although at a lower significance level (p=0.071–0.093). We also demonstrate that significantly better performance than expert agreement can be achieved if a classifier with high recall or precision was required.

Introduction

Global trade relies on the international shipping industry, which has been implicated in the spread of many marine non-indigenous species (NIS) around the world[1,2]. Modern vessels have two primary pathways for translocating NIS, namely (i) as stowaways in ballast water, or (ii) attached to the vessel surface as biofouling[3]; examples of each follow. Ballast water was the likely vector for zebra mussels to spread from Europe to the great lakes in North America[4], where they have led to increases in toxic blue-green algae[5] and cost industry more than \$200 million per year in maintaining water intake structures[6]. Biofouling is one of the most significant pathways for the spread of non-indigenous seaweeds[7–9], which can outcompete native species[10], make native kelp forests less resilient[11] and adversely impact fishing and tourism operations[12,13].

Although vessels are incentivised to manage their biofouling to reduce hydrodynamic drag and fuel costs[3,14], it is a challenging undertaking and can occur even on hulls that employ current best practice[15,16]. The primary method of biofouling management is the regular application of anti-fouling coatings. These contain biocides, such as copper, or create a surface that releases organisms or dissuades attachment to slow down the process of biofouling accumulation[17]. A vessel's operating profile contributes to fouling risk, with extended periods of inactivity being associated with higher biofouling pressure[18]. Niche areas, such as sea chests, propellers, and other complex surface structures are at high risk of becoming fouled as they can offer a sheltered environment for fouling organisms to establish. They are also a lower priority for management as they contribute less to hydrodynamic drag compared to the flat surfaces of the hull[19].

There is growing interest in closer management of the biofouling pathway by biosecurity regulators[20,21]. New Zealand has implemented a clean hull standard that sets requirements for vessels to manage biofouling and proposed a clean hull threshold to determine the potential biosecurity risk of a vessel[22]. For vessels staying longer than three weeks or visiting areas other than those designated as places of first arrival, any macrofouling other than goose barnacles is considered to be a biosecurity risk, while for short stay vessels there are macrofouling coverage thresholds[16,23]. In implementing this policy, they have stressed the vessel management requirements rather than the thresholds, as even with best management practices ships can become fouled[16].

1

Australia has also proposed requirements for vessels to implement biofouling management practices or provide evidence that their fouling is appropriately controlled[21].

In-water inspections are the best way to verify biofouling standards are being met and to collect the necessary data to measure the effectiveness of biofouling management practices. However, in-water inspections are expensive, require specialist dive teams to operate in an environment with a number of health and safety risks, and while inspections are being conducted vessels are restricted in the activities they can undertake[24]. A biofouling expert also either needs to be present during the inspection, or review the images and footage gathered afterwards, which can be a costly and time-consuming process. An alternative is to employ an underwater drone or remotely operated vehicle (ROV), which would enhance data collection opportunities but also potentially increase the burden of the expert interpreting the data.

In this paper we explore the potential for deep learning, a type of machine learning which models phenomena using deep neural networks, to automate or assist the analysis of biofouling inspection data. In the last decade deep learning has revolutionised computer vision; in fact, many regard AlexNet, the 2012 winner of the ImageNet visual recognition challenge[25], as the watershed moment for deep learning[26]. AlexNet was among the first deep convolutional neural networks (CNN)[27], an architecture that is particularly suited to computer vision tasks. Our present approach is motivated by the plethora of successful applications of deep CNNs to complex image recognition tasks, from identification of wild animals in camera trap images[28, 29] to identification of coral species[30].

A prominent example of automating biofouling image analysis is CoralNet, a machine learning method initially designed for annotating benthic surveys of coral reefs using a random annotation point approach[31]. CoralNet has been applied to assess the level of cover of different species and higher level taxonomic groups present in fouling communities on oil platforms in the UK continental shelf, using images taken by ROVs[32]. Our aim in this current study was to develop a method that could be used to assess biosecurity risk, and in this context CoralNet was less suitable. Most of the images of vessel hulls that were available for developing our method had limited biofouling coverage. Unlike coral reefs and oil platforms, vessels are not stationary and actively manage their biofouling. This makes sampling error an important consideration for annotation point approaches, like CoralNet, and as CNNs consider the whole image they do not have this weakness.

Determining the potential biosecurity risk of a vessel also does not require the identification of particular species. It has been found that there is a positive relationship between the degree of biofouling present on a vessel and the number of NIS present[23]. This has led many jurisdictions, such as New Zealand, to require biofouling to be managed holistically rather than targeting specific species[16]. Species-based approaches also scale poorly in the marine context, as there is a large number of species that can be observed in these communities, the taxonomy is highly complex, and previously unobserved species are common[20, 21]. Instead, we aimed to identify the presence and severity of biofouling which is a much simpler problem. This also makes our model more transferable to locations outside of the support of the current data and it may be of use to other jurisdictions, although performance will likely be improved by the introduction of local examples to the training data.

Methods

Dataset

We assembled a dataset of 10,263 images collected from in-water surveys of around 300 commercial and recreational vessels. This dataset comprised images provided by three jurisdictions, namely: the Australian Department of Agriculture, Water and the Environment (DAWE), the New Zealand Ministry for Primary Industries (MPI), and the California State Lands Commission (CSLC). Examples from the CSLC dataset are available in the literature[33], and the MPI data set has previously been used to inform vessel biofouling management in New Zealand[16, 23, 34].

Each image was accompanied by a label according to the six-class Level of Fouling (LoF) scheme[35]. However, due to inconsistencies in the LoF labelling across the three jurisdictions, it was necessary to relabel the dataset in a systematic way. We first devised a Simplified Level of Fouling (SLoF) scale (Table 1). This scale was based on the LoF scheme, but collapsed the six levels into pairs to create a three-class scale. This was the simplest possible set of annotations that supported our goal of identifying images with fouling present and highlighting images with severe fouling.

Table 1. Simplified Level of Fouling (SLoF) scale

Rank	Description
0	No fouling organisms, but biofilm or slime may be present.
1	Fouling organisms (e.g. barnacles, mussels, seaweed or tubeworms) are visible but patchy (1-15% of surface covered).
2	A large number of fouling organisms are present (16-100% of surface covered).

(a) SLoF 0   **(b)** SLoF 1   **(c)** SLoF 2
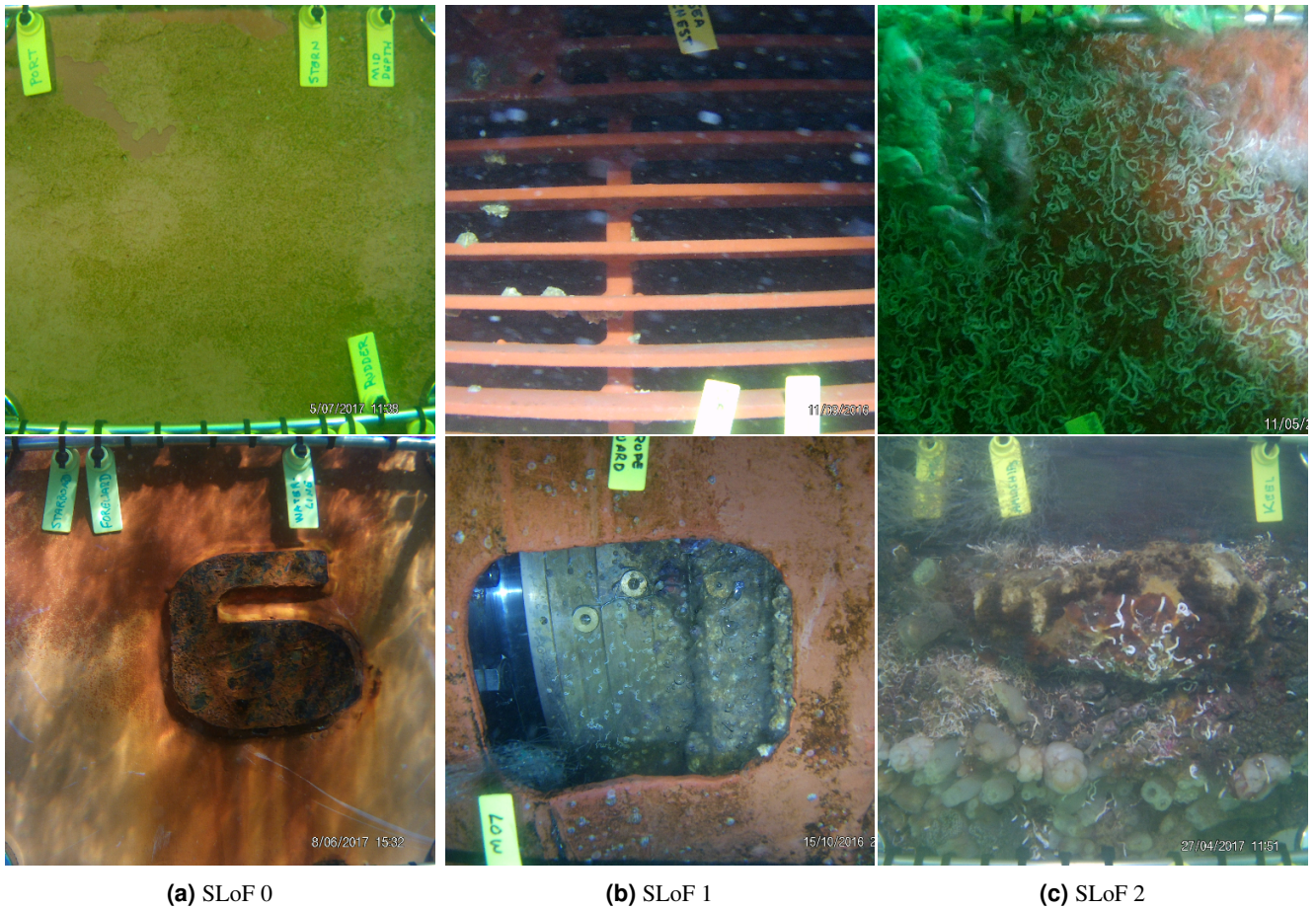
Figure 1. Example images of different SLoF levels.

We then separately asked experts and workers from the Amazon Mechanical Turk platform to annotate images according to our SLoF scheme. For the former, we engaged three biofouling experts from Ramboll New Zealand who hold qualifications in marine biology. Due to time and budget constraints, we only asked the experts to grade a set of 120 images from the DAWE dataset, constructed by stratified random sampling to ensure balance across LoF. We will call this set of 120 images the expert-labelled dataset.

Next, the convenience of the Amazon Mechanical Turk platform enabled us to annotate each image in our 10,263 dataset according to the new SLoF scheme. The examples and user interface supplied to workers are given in the supporting information. Nine workers graded each image, and the results were aggregated by taking the median value. This labelling scheme produced an imbalanced dataset, with around 70% of the images being labelled SLoF 0 compared to 20% SLoF 1 and 10% SLoF 2 (Table 2). Example images and their SLoF labels are provided in Figure 1.

Table 2. Breakdown of the number of images by SLoF in the five-fold cross validation and test datasets.

	Crossvalidation dataset	Test dataset	Total
Images with SLoF 0	6418	410	6828
Images with SLoF 1	2086	224	2310
Images with SLoF 2	918	207	1125
Total	9422	841	10263

Finally, we divided the overall dataset of 10,263 images into a training set and a test set, as is commonly done in machine learning to enable proper evaluation. The test set consists of the 120 expert-labelled images plus 721 other images from 14 vessels selected with varying degrees of fouling as determined by SLoF. The test set was constructed to challenge the machine learning model with different styles of vessel niches and fouling communities. Hence, we had a total of 841 images in the test

set; the remaining data were used to both train the deep learning model and perform cross-validation (5-fold) for hyperparameter tuning.

Machine Learning

A machine learning algorithm typically learns by training on a set of examples. We present to the machine learning algorithm a set of images with the accompanying SLoF labels (i.e. 0, 1, 2). We wish the algorithm to accurately label images *outside* of this training set, i.e., to generalize to never-before-seen images.

The setup so far makes the problem a classic supervised learning task. However unlike most image classification problems, our classes are ordinal. For example, mistaking an image of SLoF 2 as 0 is a larger error compared to mistaking an image of SLoF 1 as 0. This is an analogous challenge to the recent APTOS 2019 Blindness Detection Kaggle competition[36], which asked participants to label the severity of a disease in images on an integer scale. Many of the best performing Kaggle entries used regression losses rather than classification losses, and we follow the same approach here as this allows the relative magnitude of errors to be captured.

To measure the model performance, we consider our three-class problem as two separate binary classification tasks: 1) identify fouled images (SLoF $= 0$ versus SLoF > 0) and 2) identify heavily fouled images (SLoF $= 2$ versus SLoF < 2). This allows us to measure the effectiveness of our model as a classifier without choosing arbitrary class thresholds. Instead of the more commonly used receiver operating characteristic (ROC) curve, we use the average precision metric because it provides a better indication of classifier performance in the case that classes are imbalanced[37, 38]. We apply the average precision metric to each of the two binary classification tasks, and report finally their average as an overall indicator of performance.

Given that we are working with image data, the natural deep learning architecture to use is the convolutional neural network (CNN). A CNN comprises an input layer, which in our case is an RGB image, and an output layer, which is a raw number that relates to the SLoF class of the image. Between these are multiple hidden layers, which are connected in a sequence and make up the architecture of network. Each layer performs an operation on the previous layer, such as convolutions, pooling operations, or matrix-matrix multiplications, and the nature of these operations are determined by trainable weights[26]. The creators of AlexNet were the first to discover that stacking a large number of these layers greatly improved performance the performance of CNNs on image-recognition tasks[27].

Training a CNN consists of many components including the selection of a network architecture, a method of optimising the weights of the network (optimiser), a differentiable function that describes network performance with different configurations of weights (loss function), optimiser parameters, an image augmentation pipeline and a learning rate schedule that modifies the size of each weight update over each *epoch* (*i.e.*, iteration through the training data). Together these components affect the quality of the trained neural network. Often the term *hyperparameter* is used to refer to parameters of the optimiser, the learning scheduler, etc. The number of possible combination of these design components is incredibly large, and the available search space for determining the best combination is limited by the amount of computing power available.

We trained and tested our deep learning models with `pytorch`[39], an open-source deep-learning library developed by Facebook. We begin the model building process by conducting a learning rate test[40], using stochastic gradient descent (SGD) as the optimiser and a default set of optimiser parameters picked from the APTOS challenge. The result of this test was used to inform a quasi-random search for the best optimiser parameters[41], drawing parameters from a Sobol sequence[42] to provide more even coverage of the search space compared to random sampling. This was done by training the model for a small number of epochs, and the best sets of optimiser parameters were chosen for further exploration in addition to the default set.

We then tested performance for different combinations of the training components. We considered mean squared error and smooth-L1 loss, which we weighted by class frequency to remove the bias introduced by the imbalance of the dataset[43]. In addition to the standard optimisation algorithm SGD, other optimisers such as Adaptive Moment Estimation (Adam)[44], Rectified Adam (RAdam)[45] and Adam with a corrected weight decay algorithm (AdamW)[46] were tested. Several learning rate schedules were examined including a multi-step learning rate decay schedule, one-cycle[47] and cosine annealing[48]. In CNNs, image augmentation pipelines are important for preventing overfitting to the training data, and two different approaches with varying complexity were tried from the APTOS competition. These applied operations to our training data images that did not change their class such as rotations, random cropping, and adjusting the colour and contrast.

We considered off-the-shelf network architectures, starting with the small `resnet18` residual network which was used to test every possible combination of the training components above. The residual network was introduced to address the vanishing gradient problem in networks with large numbers of layers by allowing inputs to skip layers, and obtained first place in the 2015 ImageNet classification challenge[49]. Once the best training components were identified we trained larger and more modern network architectures on larger images, allowing us to determine if increasing image size from 224×224 to 448×448 pixels improved performance. These architectures included the "ResNeXT" squeeze and excitation networks which built upon the residual learning idea and introduced a squeeze and excitation block that incorporates relationships between image colour channels[50, 51]. We also tested the inception architecture, which attempts to identify features at different scales in

the image by applying convolution layers with several different sized kernels simultaneously[52]. We also considered efficient nets, which incorporate some of these previous ideas into an architecture that is designed to scale optimally and efficiently when the number of layers is increased[53]. A summary of the network architectures used in this paper and the `Python` packages used to implement them are provided in Table 3.

Table 3. Summary of neural network architectures used in model building.

Network Family	Network Architecture	Source package	Reference	Layers	Trainable weights (10^6)
Residual learning	`resnet18`	`torchvision`	[49]	18	11
Squeeze and excitation	`se_resnext50_32x4d`	`pretrainedmodels`	[50,51]	50	25
Squeeze and excitation	`se_resnext101_32x4d`	`pretrainedmodels`	[50,51]	101	47
Inception	`inceptionv4`	`pretrainedmodels`	[52]	150	41
Inception	`inceptionresnetv2`	`pretrainedmodels`	[52]	245	54
Efficient-net	`efficientnet-b3`	`efficientnet-pytorch`	[53]	27	11
Efficient-net	`efficientnet-b4`	`efficientnet-pytorch`	[53]	33	17
Efficient-net	`efficientnet-b5`	`efficientnet-pytorch`	[53]	39	28

We used the pretrained ImageNet weights to initialise all our networks. These weights are obtained by pre-training the network on the ImageNet database, which contains millions of images with a thousand different categories[54], and were available for download through the architecture packages. This is a common practice known as transfer learning which reduces the number of epochs required to reach a performance plateau and improves results on small datasets[28]. All network weights were trained, except for the batch-normalisation layers, as these are best trained on large datasets like the ImageNet database.

The final step was creating a network ensemble. This is a technique where the class of an image is predicted by multiple networks, and their outputs are combined to obtain better performance[28]. We took the simplest approach, which is to average the raw network output. We identified the best performing ensemble by testing the performance of every combination of network trained on a particular image size. This gave us 510 possible ensembles to test for each image resolution. The full details of the model fitting process are provided in the supporting documentation.

Thresholding to create a classifier
The raw output of our model is a single number which needs to be thresholded to map back to the SLoF classes. The precision-recall curve created by combining validation crossfolds is used to guide this mapping process (Figure 2). In particular, the curve highlights the trade-off between precision and recall when choosing a threshold. A high precision classifier will only capture some of the positive results, while a high recall classifier will capture most of the positive results along with many false positives. For illustrative purposes we have selected three classifiers to explore, namely a *high-precision classifier* chosen with a 25% recall threshold, a *high-recall classifier* chosen with a 90% recall threshold and a *balanced classifier* with a 70% recall threshold.

Comparison to experts
Perfect agreement within our labelling scheme is unlikely among biofouling experts due to its subjectivity, and the frequency at which experts agree with each other is a useful benchmark to evaluate the performance of our alternative labelling methods. Our 120-image expert-labelled dataset was graded by three experts, yielding a total of 720 expert-expert label pairs. These were obtained by pairing the labels of one expert to annotations provided by the other two, and repeating the process for each expert. We also paired the MTurk and model labels with each expert, providing 360 label pairs to compare to the expert-expert label pair performance.

We assessed the significance of differences in precision and recall with a two-sided Fisher's exact test[55] with the `fisher.test` function in R[56], using the null hypothesis that the precision or recall between experts is no different to the precision or recall with our alternate labels and using experts as the ground truth. We also used the two-one-sided t-tests (TOST) approach to test for non-inferiority[57] using the `TOSTER` R package[58]. The null hypothesis in this method was that the agreement observed between experts would be at least 5% better compared to the agreement observed between our alternate labels and experts. A separate non-inferiority test was necessary as the lack of significant differences does not mean we can conclude that two distributions are similar[59]. We chose a p-value of 0.05 to signify statistical significance.

Results

Model thresholding and performance
Our best performing model based on five fold cross validation was an ensemble consisting of the `resnet18`, `se_resnext50_32x4d`, `inceptionv4`, `inceptionresnetv2`, `efficientnet-b3` and `efficientnet-b5` CNN architectures (see Table
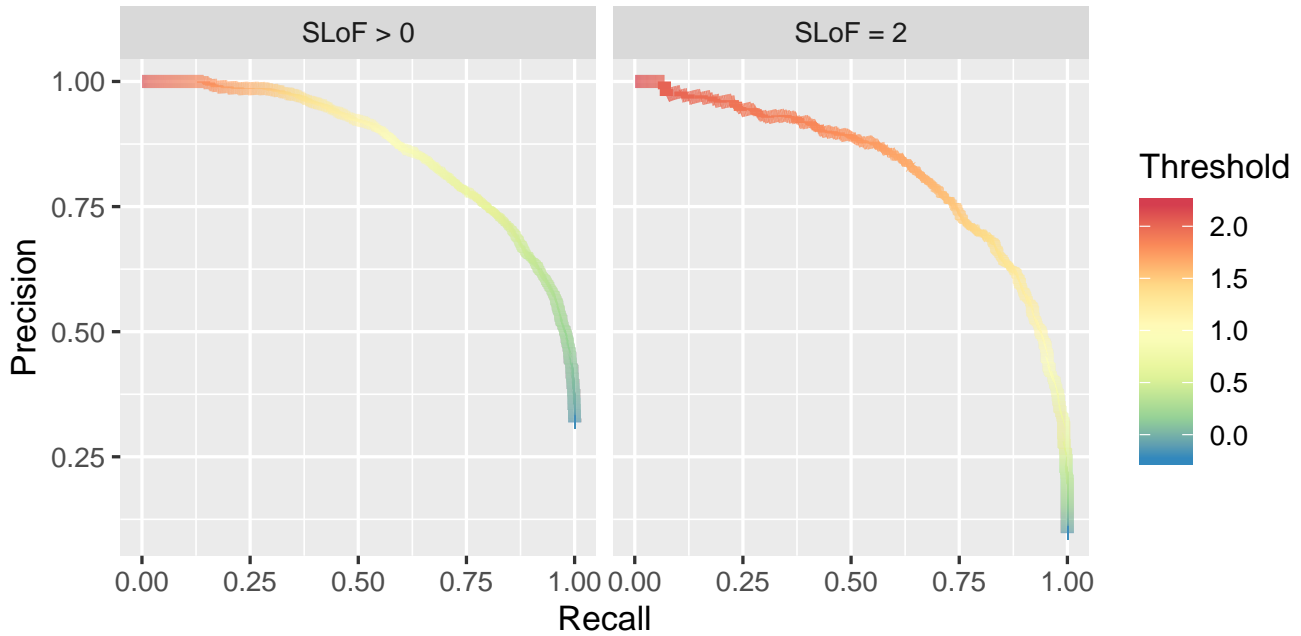
Figure 2. Precision-recall curve for model using validation data from each crossfold.
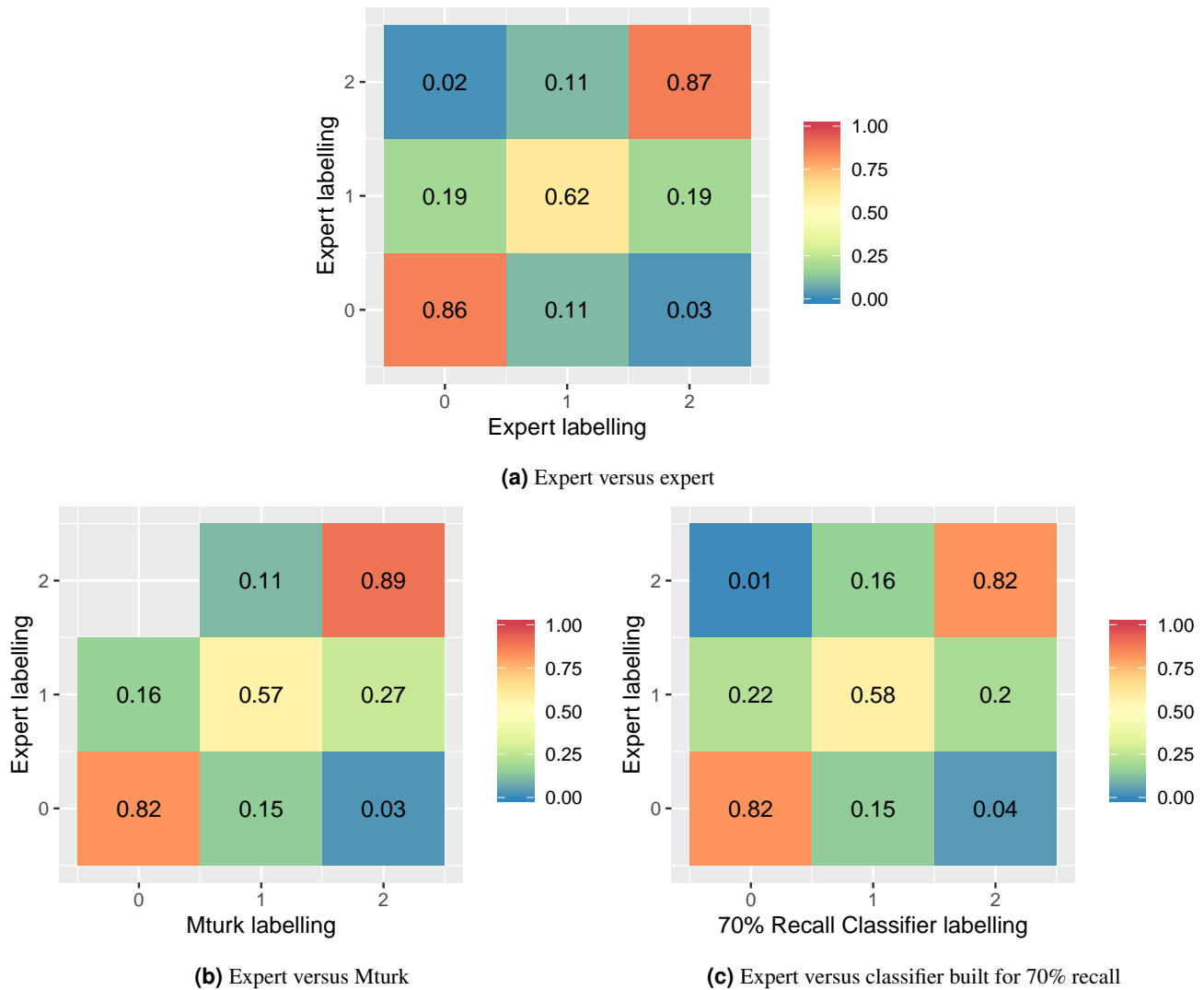
3) with an input image size of 448×448 pixels, trained using a SGD optimiser, the default set of optimiser hyperparameters, smooth-L1 loss, a cosine annealing learning rate schedule and the more complex set of image augmentations. This gave a mean average precision of 0.849 (standard deviation of 0.018), which significantly improved upon the results from our initial random search of 0.799 (standard deviation of 0.028). The full results of the model fitting process is provided in the supporting documentation. The results for each binary classification problem with this model on our validation data and test dataset are shown in Table 4. The classifiers show better results on the testing dataset, which is promising for the generalisability of our model.

Table 4. Precision and recall of classifier using model with chosen recall thresholds on the validation and testing dataset.

Data	SLoF	Threshold	Precision	Recall
Validation	>=1	0.426	0.646	0.900
Test	>=1	0.426	0.796	0.961
Validation	>=1	0.782	0.817	0.700
Test	>=1	0.782	0.914	0.810
Validation	>=2	1.218	0.581	0.900
Test	>=2	1.218	0.742	0.918
Validation	>=2	1.577	0.780	0.700
Test	>=2	1.577	0.939	0.744
Validation	>=1	1.623	0.986	0.250
Test	>=1	1.623	1.000	0.350
Validation	>=2	1.918	0.943	0.251
Test	>=2	1.918	1.000	0.357

Inter-rater reliability

We found that experts agree most often on images showing clean or heavily fouled hulls, while images that only contained some fouling was were more likely to obtain inconsistent grades (Figure 3a). Overall, experts showed 89% agreement for both tasks (95% CI: 87–92%). As we have considered every combination of experts, the recall and precision calculated for each task was the same. Experts were found to achieve 91% precision and recall for identifying images containing fouling (95% CI: 88–94%) and 87% for images containing heavy fouling (95% CI: 82–90%) (Table 5).

(a) Expert versus expert

(b) Expert versus Mturk

(c) Expert versus classifier built for 70% recall

Figure 3. Confusion matrices for expert versus MTurk and model labels using the SLoF score on the 120 image test set.

When the rate of agreement between the MTurk labels and experts was compared to agreement between experts we found that the difference was not statistically significant when comparing recall (p = 0.28–0.44) or precision (p = 0.32–0.48) for both tasks. However, the non-inferiority test showed that the agreement of the MTurk labels to experts was similar to within a margin of at most 5% worse (p=0.004–0.02) This similarity could also be observed from the confusion matrix between expert and MTurk labelling (Figure 3b).

Expert agreement is a useful benchmark for our computer vision model, and depending on the thresholds chosen to create a classifier, different outcomes were found (Table 5). Choosing a 70% recall threshold for both tasks resulted in a classifier that was close to having similar agreement with experts to within a margin of at most 5% worse (p = 0.071–0.093). While these p-values were slightly above our threshold for statistical significance, it was also noted that the difference in precision (p = 0.33–0.65) or recall (p = 0.25–0.88) was also not statistically significant. The results for this classifier are shown in Figure 3c as a confusion matrix. Using a 90% recall threshold instead could produce significantly higher recall or close to with respect to experts (p = 0.019–0.058) at the cost of significantly lower precision (p=0.002–0.004). Conversely, using a much lower recall threshold of 25% results in significantly higher precision (p=0.004–0.018), with a corresponding decrease in recall (p< 0.001).

Discussion

In this study we applied deep learning methods to identify the presence and severity of biofouling on ship hulls using images annotated on the Amazon Mechanical Turk crowdsourcing platform, and compared our performance to experts. Our MTurk

Table 5. Precision and recall for expert-expert, MTurk-expert and classifier-expert label pairs. Numbers in brackets are the 95% confidence intervals. The TOST column are non-inferiority testing p-values using the two-one-sided t-tests approach, with the null hypothesis being that the agreement observed between experts would be at least 5% better compared to the agreement observed for the method-expert label pairs. The p-value columns are given by a two-sided exact Fisher test, with the null hypothesis being that the method-expert label pairs do not differ in their precision or recall compared to the expert-expert label pairs.

Labels	Classification	Label pairs	Agreement	TOST	Recall	p-value	Precision	p-value
Expert	$SLoF > 0$	720	0.89 (0.87-0.92)		0.91 (0.88-0.94)		0.91 (0.88-0.94)	
Mturk	$SLoF > 0$	360	0.90 (0.86-0.93)	0.004	0.94 (0.90-0.97)	0.279	0.90 (0.85-0.93)	0.484
90% Recall Classifier	$SLoF > 0$	360	0.86 (0.81-0.89)	0.305	0.96 (0.92-0.98)	0.058	0.84 (0.78-0.88)	0.002
70% Recall Classifier	$SLoF > 0$	360	0.88 (0.84-0.91)	0.071	0.91 (0.86-0.94)	0.884	0.89 (0.84-0.93)	0.328
25% Recall Classifier	$SLoF > 0$	360	0.71 (0.66-0.75)	1.000	0.54 (0.47-0.60)	< 0.001	0.98 (0.93-0.99)	0.018
Expert	$SLoF = 2$	720	0.89 (0.87-0.92)		0.87 (0.82-0.90)		0.87 (0.82-0.90)	
Mturk	$SLoF = 2$	360	0.89 (0.85-0.92)	0.020	0.89 (0.83-0.94)	0.440	0.83 (0.76-0.89)	0.323
90% Recall Classifier	$SLoF = 2$	360	0.86 (0.82-0.89)	0.261	0.94 (0.89-0.98)	0.019	0.76 (0.69-0.82)	0.004
70% Recall Classifier	$SLoF = 2$	360	0.87 (0.83-0.90)	0.093	0.82 (0.75-0.88)	0.250	0.85 (0.78-0.90)	0.654
25% Recall Classifier	$SLoF = 2$	360	0.78 (0.73-0.82)	0.997	0.44 (0.35-0.52)	< 0.001	0.98 (0.91-1.00)	0.004

labels showed similar agreement to experts, and this result is highly promising as it suggests images can be effectively graded for the presence and severity of biofouling by non-experts by aggregating their annotations, offering the potential for substantial time and cost-savings. This labelling was also sufficient for training CNNs that were found to have close to expert agreement, although fine-tuning our model on expert-annotated data may offer further improvements. We have also demonstrated that if high precision or recall is desired for the application of the model, then classifiers can be created that offer better performance than experts with regard to this property. This allows the behaviour of the classifiers to be tuned for a particular application. For example, when screening vessels for biosecurity risk it may be desirable to have a classifier with higher recall so few images with severe fouling are missed. Conversely, if an activity were being undertaken where intervention capacity was limited then a classifier with higher precision would be more appropriate.

The effectiveness of management activities for vessel biofouling in reducing biosecurity risk is currently a key knowledge gap for regulators, which makes it difficult to determine which combination of activities will provide confidence that a vessel is low risk. This model could be applied to provide a cheaper and more reliable way to identify the most effective management strategies, if combined with standardised vessel sampling protocols[24, 60], clear definitions of vessel biosecurity risk, such as the clean hull standard for New Zealand[16], collection of management data, and ongoing in-water vessel inspections. This will also support more consistent assessment of effective management strategies between different organisations, which is a limitation of expert assessments. Building this evidence base would also provide benefits to industry, as it would be a basis from which to work towards regulatory alignment between different jurisdictions.

In-water cleaning and hull grooming are increasingly important biofouling management activities, as regular cleaning can limit biofouling accumulation and provide options where the anti-fouling coatings of vessels are no longer effective or have failed[61, 62]. However, it also presents a biosecurity risk because cleaning can lead to the release of viable propagules and organisms can detach and still be viable[63–65]. One way this risk can be managed is by considering the biofouling state of the vessel before setting conditions on in-water cleaning or grooming activities. For example, New Zealand recommends that in-water cleaning of macrofouling with an international origin must capture biological waste and dispose of it on land or be rendered non-viable, but this would not be necessary if only a slime layer were present[66]. Automatic detection of biofouling using the state-of-the-art deep learning tools developed in this paper could be a cost-effective and reliable way for regulators and industry to process the outcomes of biofouling inspections for this purpose.

So far we have only tested our model on static images. Since videos are constructed using a stream of images, our model should be readily adaptable to videos as well. However, further work is needed to address issues such as identifying the frames in which the camera is directed towards a vessel hull as opposed to open water or where image quality is poor, which is a common issue when analysing stills obtained from ROV footage[32]. The video format would also offer the opportunity to incorporate information from future and previous frames to improve and smooth fouling estimates, and ideas from current action recognition methods could potentially be applied[67].

Our SLoF labelling scheme also only relates to the percentage cover of macrofouling present within an image, which could be more rigorously used to determine the absolute biosecurity risk of a vessel if the area of hull captured within the image could be estimated. Given that in-water inspection methods are expected to vary greatly between jurisdictions, being able to do this without the presence of scale bars would be a major advantage. One possibility would be training a deep neural network on images of vessel hulls taken using multiple cameras, and building a model that will estimate depth given a single image[68].

References

1. Hayes, K. R. & Sliwa, C. Identifying potential marine pests—a deductive approach applied to Australia. *Mar. Pollut. Bull.* **46**, 91–98 (2003).

2. Clarke, S., Hollings, T., Liu, N., Hood, G. & Robinson, A. Biosecurity risk factors presented by international vessels: a statistical analysis. *Biol. Invasions* **19**, 2837–2850 (2017).

3. Floerl, O. & Coutts, A. Potential ramifications of the global economic crisis on human-mediated dispersal of marine non-indigenous species. *Mar. Pollut. Bull.* **58**, 1595–1598 (2009).

4. Johnson, L. E. & Carlton, J. T. Post-establishment spread in large-scale invasions: dispersal mechanisms of the zebra mussel *Dreissena polymorpha*. *Ecology* **77**, 1686–1690 (1996).

5. Vanderploeg, H. A. *et al.* Zebra mussel (*Dreissena polymorpha*) selective filtration promoted toxic *Microcystis blooms* in Saginaw Bay (Lake Huron) and Lake Erie. *Can. J. Fish. Aquatic Sci.* **58**, 1208–1221 (2001).

6. Connelly, N. A., O'Neill, C. R., Knuth, B. A. & Brown, T. L. Economic impacts of zebra mussels on drinking water treatment and electric power generation facilities. *Environ. management* **40**, 105–112 (2007).

7. Dürr, S. & Thomason, J. C. *Biofouling* (John Wiley & Sons, 2009).

8. Farrell, P. & Fletcher, R. Boats as a vector for the introduction and spread of a fouling alga, *Undaria pinnatifida* in the UK. *Porcup. Mar. Nat. Hist. Soc. Newsl.* (2004).

9. Williams, S. L. & Smith, J. E. A global review of the distribution, taxonomy, and impacts of introduced seaweeds. *Annu. Rev. Ecol. Evol. Syst.* **38**, 327–359 (2007).

10. Levin, P. S., Coyer, J. A., Petrik, R. & Good, T. P. Community-wide effects of nonindigenous species on temperate rocky reefs. *Ecology* **83**, 3182–3193 (2002).

11. Bulleri, F., Benedetti-Cecchi, L., Ceccherelli, G. & Tamburello, L. A few is enough: a low cover of a non-native seaweed reduces the resilience of Mediterranean macroalgal stands to disturbances of varying extent. *Biol. Invasions* **19**, 2291–2305 (2017).

12. Freshwater, D. W. *et al.* Distribution and identification of an invasive *Gracilaria* species that is hampering commercial fishing operations in southeastern North Carolina, USA. *Biol. invasions* **8**, 631–637 (2006).

13. Meretta, P. E., Matula, C. V., Casas, G. *et al.* Occurrence of the alien kelp *Undaria pinnatifida* (Laminariales, Phaeophyceae) in Mar del Plata, Argentina. *BioInvasions Rec.* **1**, 59–63 (2012).

14. Schultz, M., Bendick, J., Holm, E. & Hertel, W. Economic impact of biofouling on a naval surface ship. *Biofouling* **27**, 87–98 (2011).

15. Davidson, I. C. *et al.* Recreational boats as potential vectors of marine organisms at an invasion hotspot. *Aquatic Biol.* **11**, 179–191 (2010).

16. Georgiades, E. & Kluza, D. Evidence-based decision making to underpin the thresholds in New Zealand's craft risk management standard: biofouling on vessels arriving to New Zealand. *Mar. Technol. Soc. J.* **51**, 76–88 (2017).

17. Chambers, L. D., Stokes, K. R., Walsh, F. C. & Wood, R. J. Modern approaches to marine antifouling coatings. *Surf. Coatings Technol.* **201**, 3642–3652 (2006).

18. Scardino, A., Fletcher, L. & Lewis, J. A. Fouling control using air bubble curtains: protection for stationary vessels. *J. Mar. Eng. & Technol.* **8**, 3–10 (2009).

19. Moser, C. S. *et al.* Quantifying the extent of niche areas in the global fleet of commercial ships: the potential for "super-hot spots" of biofouling. *Biol. Invasions* **19**, 1745–1759 (2017).

20. Davidson, I. *et al.* Mini-review: Assessing the drivers of ship biofouling management–aligning industry and biosecurity goals. *Biofouling* **32**, 411–428 (2016).

21. Hayes, K. R., Inglis, G. & Barry, S. C. The assessment and management of marine pest risks posed by shipping: The Australian and New Zealand experience. *Front. Mar. Sci.* **6**, 489 (2019).

22. Ministry for Primary Industries. Biofouling on vessels arriving to new zealand. New Zealand Government (2018).

23. Georgiades, E. *et al.* Regulating vessel biofouling to support New Zealand's marine biosecurity system – A blue print for evidence-based decision making. *Front. Mar. Sci.* **7**, 390 (2020).

24. Zabin, C. *et al.* How will vessels be inspected to meet emerging biofouling regulations for the prevention of marine invasions? *Manag. Biol. Invasions* **9**, 195–208 (2018).

25. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

26. Rawat, W. & Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* **29**, 2352–2449 (2017).

27. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105 (2012).

28. Norouzzadeh, M. S. *et al.* Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci.* **115**, E5716–E5725 (2018).

29. Tabak, M. A. *et al.* Machine learning to classify animal species in camera trap images: applications in ecology. *Methods Ecol. Evol.* **10**, 585–590 (2019).

30. Gómez-Ríos, A., Tabik, S., Luengo, J., Shihavuddin, A. & Herrera, F. Coral species identification with texture or structure images using a two-level classifier based on Convolutional Neural Networks. *Knowledge-Based Syst.* **184**, 104891 (2019).

31. Beijbom, O., Edmunds, P. J., Kline, D. I., Mitchell, B. G. & Kriegman, D. Automated annotation of coral reef survey images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1170–1177 (IEEE, 2012).

32. Gormley, K. *et al.* Automated image analysis of offshore infrastructure marine biofouling. *J. Mar. Sci. Eng.* **6**, 2 (2018).

33. Davidson, I. C., Scianni, C., Minton, M. S. & Ruiz, G. M. A history of ship specialization and consequences for marine invasions, management and policy. *J. Appl. Ecol.* **55**, 1799–1811 (2018).

34. Bell, A., Phillips, S., Georgiades, E., Kluza, D. & Denny, C. Risk analysis: Vessel biofouling. *Minist. Agric. For. Wellington, New Zealand* (2011).

35. Floerl, O., Inglis, G. J. & Hayden, B. J. A risk-based predictive tool to prevent accidental introductions of nonindigenous marine species. *Environ. Manag.* **35**, 765–778 (2005).

36. Kaggle. APTOS 2019 Blindness Detection. https://www.kaggle.com/c/aptos2019-blindness-detection (2020). Accessed: 2020-05-07.

37. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10** (2015).

38. Liu, Z. & Bondell, H. D. Binormal precision–recall curves for optimal classification of imbalanced data. *Stat. Biosci.* **11**, 141–161 (2019).

39. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

40. Smith, L. N. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 464–472 (IEEE, 2017).

41. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. machine learning research* **13**, 281–305 (2012).

42. Atanassov, E. I. A new efficient algorithm for generating the scrambled Sobol sequence. In *International Conference on Numerical Methods and Applications*, 83–90 (Springer, 2002).

43. Yue, S. Imbalanced malware images classification: a CNN based approach. *arXiv preprint arXiv:1708.08042* (2017).

44. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

45. Liu, L. *et al.* On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265* (2019).

46. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

47. Smith, L. N. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820* (2018).

48. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

49. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

50. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500 (2017).

51. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).

52. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence* (2017).

53. Tan, M. & Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).

54. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *Int. journal computer vision* **115**, 211–252 (2015).

55. Fisher, R. A. Statistical methods for research workers. In *Breakthroughs in statistics*, 66–70 (Springer, 1992).

56. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019).

57. Walker, E. & Nowacki, A. S. Understanding equivalence and noninferiority testing. *J. general internal medicine* **26**, 192–196 (2011).

58. Lakens, D. Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Soc. Psychol. Pers. Sci.* **1**, 1–8, DOI: 10.1177/1948550617697177 (2017).

59. Blackwelder, W. C. "Proving the null hypothesis" in clinical trials. *Control. clinical trials* **3**, 345–353 (1982).

60. Georgiades, E. & Kluza, D. Technical advice: Conduct of in-water biofouling surveys for domestic vessels. https://www.mpi.govt.nz/dmsdocument/40424-technical-advice-conduct-of-in-water-biofouling-surveys-for-domestic-vessels (2020). Accessed: 2020-08-20.

61. Hunsucker, K. Z., Ralston, E., Gardner, H. & Swain, G. Specialized grooming as a mechanical method to prevent marine invasive species recruitment and transport on ship hulls. In *Impacts of Invasive Species on Coastal Environments*, 247–265 (Springer, 2019).

62. Tribou, M. & Swain, G. The effects of grooming on a copper ablative coating: a six year study. *Biofouling* **33**, 494–504 (2017).

63. Hopkins, G. A. & Forrest, B. M. Management options for vessel hull fouling: an overview of risks posed by in-water cleaning. *ICES J. Mar. Sci.* **65**, 811–815 (2008).

64. Tamburri, M. N. *et al.* In-water cleaning and capture to remove ship biofouling: An initial evaluation of efficacy and environmental safety. *Front. Mar. Sci.* (2020).

65. Scianni, C. & Georgiades, E. Vessel in-water cleaning or treatment: Identification of environmental risks and science needs for evidence-based decision making. *Front. Mar. Sci.* **6**, 467 (2019).

66. Georgiades, E., Growcott, A. & Kluza, D. Technical guidance on biofouling management for vessels arriving to New Zealand. *Minist. for Prim. Ind. Wellingt.* 77665–793 (2018).

67. Shi, Y., Tian, Y., Wang, Y. & Huang, T. Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Transactions on Multimed.* **19**, 1510–1520 (2017).

68. Mal, F. & Karaman, S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 1–8 (IEEE, 2018).

Acknowledgements

Data Availability

The data that support the findings of this study are available from DAWE, MPI and CSLC, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the respective organisations.

Author contributions statement

NB wrote draft manuscript, curated data, directed labelling and coded and performed all analyses. SW advised on machine learning and access to supercomputer resources. AR advised on statistical analysis and project direction. BW and PW solicited data and advised on project direction. All authors reviewed the manuscript.

Additional Information

The authors declare no competing interests.