Progress in Oceanography 196 (2021) 102612

Contents lists available at ScienceDirect

Progress in Oceanography

journal homepage: www.elsevier.com/locate/pocean

Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance

Jennifer M. Durden [a,*], Brett Hosking [a], Brian J. Bett [a], Danelle Cline [b], Henry A. Ruhl [a,b]

[a] National Oceanography Centre, Southampton, UK
[b] Monterey Bay Aquarium Research Institute, Moss Landing, USA

ARTICLE INFO

Keywords:
Computer vision
Deep learning
Benthic ecology
Image annotation
Marine photography
Artificial intelligence
Convolutional neural networks
Sample size

ABSTRACT

Machine learning is rapidly developing as a tool for gathering data from imagery and may be useful in identifying (classifying) visible specimens in large numbers of seabed photographs. Application of an automated classification workflow requires manually identified specimens to be supplied for training and validating the model. These training and validation datasets are generally generated by partitioning the available manual identified specimens; typical ratios of training to validation dataset sizes are 75:25 or 80:20. However, this approach does not facilitate the desired scalability, which would require models to successfully classify specimens in hundreds of thousands to millions of images after training on a relatively small subset of manually identified specimens. A second problem is related to the 'class imbalance', where natural community structure means that fewer specimens of rare morphotypes are available for model training. We investigated the impact of independent variation of the training and validation dataset sizes on the performance of a convolutional neural network classifier on benthic invertebrates visible in a very large set of seabed photographs captured by an autonomous underwater vehicle at the Porcupine Abyssal Plain Sustained Observatory. We tested the impact of increasing training dataset size on specimen classification in a single validation dataset, and then tested the impact of increasing validation set size, evaluating ecological metrics in addition to computer vision metrics. Computer vision metrics (recall, precision, F1-score) indicated that classification improved with increasing training dataset size. In terms of ecological metrics, the number of morphotypes recorded increased, while diversity decreased with increasing training dataset size. Variation and bias in diversity metrics decreased with increasing training dataset size. Multivariate dispersion in apparent community composition was reduced, and bias from expert-derived data declined with increasing training dataset size. In contrast, classification success and resulting ecological metrics did not differ significantly with varying validation dataset sizes. Thus, the selection of an appropriate training dataset size is key to ensuring robust automated classifications of benthic invertebrates in seabed photographs, in terms of ecological results, and validation may be conducted on a comparatively small dataset with confidence that similar results will be obtained in a larger production dataset. In addition, our results suggest that automated classification of less common morphotypes may be feasible, providing that the overall training dataset size is sufficiently large. Thus, tactics for reducing class imbalance in the training dataset may produce improvements in the resulting ecological metrics.

1. Introduction

Machine learning is developing as a tool for gathering data from imagery, and marine ecologists have begun to apply it to exploit photographic datasets (e.g. Beijbom et al., 2015; Hu & Davis, 2005; Matabos et al., 2017; Purser et al., 2009). It can be used to count and identify specimens in imagery (generally megafauna >1 cm; Bett, 2019), data critical to monitoring benthic communities to detect natural and anthropogenic change, and for conservation purposes (Danovaro et al., 2020). The basis for such monitoring is benthic ecological studies. Imagery datasets for these studies are increasing in size (number of photographs, hours of video) as camera, platform, and battery technologies

* Corresponding author.
E-mail address: jennifer.durden@noc.ac.uk (J.M. Durden).

https://doi.org/10.1016/j.pocean.2021.102612
Received 17 July 2020; Received in revised form 7 May 2021; Accepted 13 May 2021
Available online 20 May 2021

improve (Durden et al., 2016b), facilitating the examination of larger areas of seafloor. Such large spatial scale seabed photography, typically using still images captured from a moving platform (e.g. autonomous underwater vehicle, remotely operated vehicle, towed platform), has been instrumental in improving our understanding of standing stocks, spatial heterogeneity, and ecosystem function in benthic environments (e.g. Benoist et al., 2019; Durden et al., 2020b, this Collection; Mitchell et al., 2020, this Collection; Morris et al., 2016; Simon-Lledó et al., 2019b). Machine learning tools may be employed to reduce the substantial effort required to manually annotate the specimens in these large datasets (e.g. Durden et al., 2016a; MacLeod et al., 2010). However, the cost of the manual effort to generate annotations for the training and validation of machine learning tools has hampered their development and application. Studies of the applications of machine learning tools to seabed images have been confined to relatively small photographic datasets (hundreds to thousands of images; e.g. Piechaud et al., 2019; Schoening et al., 2012), with subsequent extrapolation of results to the dataset sizes possible with contemporary seabed photographic survey techniques (hundreds of thousands to millions of images).

An automated annotation workflow to generate ecological data from images involves the processes of detection, segmentation, and classification. Detection identifies the presence and location of any specimens in an image, segmentation identifies the pixels belonging to a specimen, and classification assigns each specimen to a category. Classification is also known as 'identification' when experts assign an organism to a morphotype or taxonomic identity. Classification is the key challenge, since state-of-the-art object detection and segmentation algorithms fundamentally rely on classification performance (e.g. He et al., 2020; Ren et al., 2017). Machine learning approaches, such as support vector machine classifiers (e.g. Beijbom et al., 2015; Schoening et al., 2012) and random forests (e.g. Osterloff et al., 2016), have been used for classification of benthic megafauna in seabed photographs. However, these approaches rely on the manual selection of visual features to train the classifier, and as a result their performance is limited by the choice of features extracted. By contrast, neural networks do not require initial feature extraction, as they learn features from the supplied training data, making them an attractive approach. Convolutional neural networks (CNN; Lecun et al., 1998) have recently been applied to the classification of fish in video (e.g. Qin et al., 2016; Salman et al., 2016) and megafauna in seabed images (e.g. Langenkämper & Nattkemper, 2016; Langenkämper et al., 2020; Piechaud et al., 2019).

The application of an automated classification workflow requires expert-identified specimens in images to be supplied for training the model, and further expert-identified specimens for use in validating the results of the model. These training and validation data are usually generated through manual annotation of images by experts, potentially representing substantial effort. Those wishing to apply a model to their collected imagery are faced with the challenge of determining the minimum expert-generated data required to adequately train and validate their model. In addition to using relatively small datasets of annotated seabed photographs, prior studies of AI for spatial ecology have largely generated the training and validation datasets by partitioning those data; ratios of training to validation dataset sizes are commonly 75:25 or 80:20 (e.g. Piechaud et al., 2019; Qin et al., 2016; Shafait et al., 2016). In effect, models are trained on small datasets and validated on even smaller ones. Cross validation may be employed as a technique to partition available annotated images and perform repeated validation, which involves small training and larger validations datasets (Kohavi, 1995), and has been used in the application of AI to identify fish in images from a fixed observatory (Marini et al., 2018); this method still requires a sufficiently-sized annotated dataset from which to draw training and validation data. However, this method of assigning data to large training and smaller validation datasets does not align well with the desired scalability of a model in real-world applications. To be effective and efficient, models must successfully classify specimens in very large image datasets after training with a relatively small subset. To accomplish such scalability, the small subset must be sufficient in size to provide adequate training data for the model, and model performance must be evaluated with much larger datasets than those that have been previously employed in model validation studies.

The natural composition of most benthic communities involves varying abundances of different taxa, and thus a substantial imbalance in the number of specimens per morphotype. This presents a problem in computer vision (known as the 'class imbalance', Langenkämper et al., 2019; Langenkämper et al., 2020), as it leads to reductions in the successful classification of less common morphotypes. Previous studies have focused on relatively common morphotypes (e.g. 1–10 morphotypes or taxa: Matabos et al., 2017; Piechaud et al., 2019; Schoening et al., 2012; Shafait et al., 2016), in effect avoiding this significant issue. However, effective real-world application of automated classification models must include less common morphotypes, and still produce ecologically robust results.

We investigated the application of automated classification of benthic invertebrate megafauna in seabed photographs, focusing on the ecological evaluation of the results. In particular, we examine the impact of variations in training and validation dataset sizes on the performance of automated classification of invertebrate megafauna visible in seabed photographs via two scenarios. Scenario 1 tests the impact of increasing training dataset size on specimen classification in a single validation dataset, with training-to-validation ratios of 80:20 to 98:2. In Scenario 2, we use a single training dataset size to test the impact of increasing validation set size. The datasets used were drawn from expert-annotated seabed photographs captured with an autonomous underwater vehicle at the Porcupine Abyssal Plain Sustained Observatory (Durden et al., 2020b, this Collection). This annotated image set is larger than those previously used in AI development for spatial ecological studies, providing a realistic assessment of the scale of annotation required to generate appropriately-sized training and validation datasets. We evaluate CNN model performance using standard computer vision metrics (recall, precision, F1-score), ecological parameters (diversity measures, community composition), and comparisons to manually generated expert data. We also consider the effects of the class imbalance in the training dataset on CNN model results.

2. Methods

2.1. Photographic and annotation data

The Porcupine Abyssal Plain Sustained Observatory (PAP-SO, 4850 m water depth; Hartman et al., 2021) is a long-term time series site where photographic data is used to assess benthic invertebrate megafaunal community dynamics (e.g. Bett et al., 2001; Durden et al., 2020a). Seabed images (2448 × 2048 pixel) were captured using the autonomous underwater vehicle Autosub6000 in the PAP-SO area, and processed for colour and illumination (Morris et al., 2014), including dark frame removal, non-uniform illumination correction, histogram correction, and steps to mosaic the imagery into strips of ten images (referred to as 'tiles'; Fig. 1). Megafaunal (>1 cm) invertebrate specimens ('samples' in computer vision nomenclature) were manually detected, their location in the image noted, and classified to the most detailed taxonomic level possible by experts. These specimen counts and identifications have been the subject of prior studies on the ecology of the PAP-SO area (e.g. Durden et al., 2020b, this Collection; Durden et al., 2017; Mitchell et al., 2020, this Collection; Morris et al., 2016), and form part of the time series of invertebrate megafauna at the site (Billett et al., 2010; Billett et al., 2001; Durden et al., 2015; Durden et al., 2020b). These data have also been used in a study comparing detection and classification performance among expert annotators (Durden et al., 2016a). Individual specimens detected by the experts in these photographs were used in this study by cropping them from the images (see Section 2.4 Specimen preparation; Fig. 1).

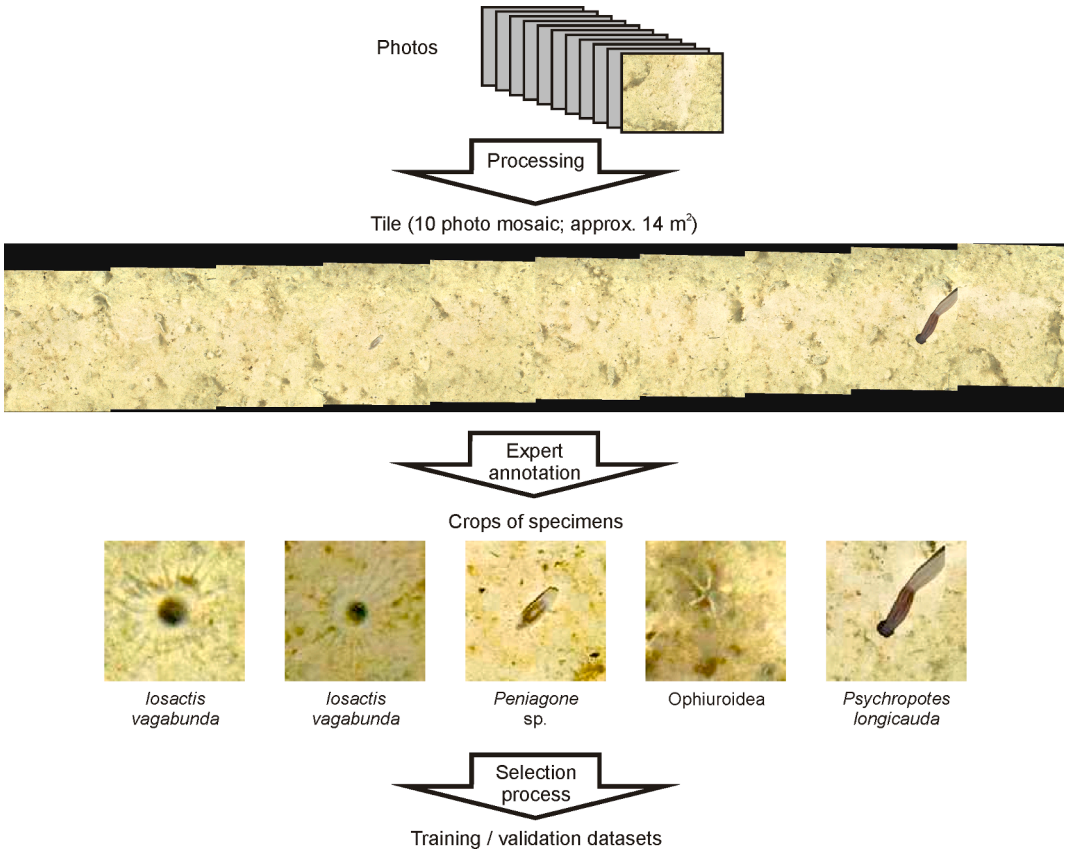

Fig. 1. Seabed photograph processing and annotation steps to prepare specimen crops for the training and validation datasets. Processing of seabed photographs to form tiles is described in Morris et al (2014).

A total of 25 benthic invertebrate megafaunal morphotypes, previously noted to be of ecological significance at the PAP-SO, were targeted for CNN classification (Table S1). These morphotypes were selected based on the following criteria: (1) only morphotypes occurring on the abyssal plain were included, and morphotypes known from only abyssal hill locations were excluded (Durden et al., 2020b), to simplify the subsequent ecological assessments; (2) only morphotypes where the expert annotations represented the most specific taxonomic identifications were used, these were generally genus- or species-level identifications; (3) only morphotypes with sufficient specimens for training and validating a model were included. The abyssal plain megafaunal community was dominated by few morphotypes: *Iosactis vagabunda* contributed 53% of specimens (see Table S1), while the least abundant morphotype included in this study contributed 0.2% of specimens. Note that the morphotype Elpidiidae spp. represents a complex of species (*Amperima rosea*, *Ellipinion mole* and *Kolga nana*) from the family Elpidiidae that are difficult to reliably distinguish in seabed photographs, while specimens of another confamilial genus, *Peniagone* sp., were considered as a distinct morphotype. A single morphotype for all ophiuroids was also used. A 'gold standard' was generated from the expert annotations for comparison to the CNN-generated classifications: all specimens detected by any expert were included, and specimens for which experts disagreed on their original classifications were re-inspected and assigned to a morphotype (as described in Durden et al., 2016a). Results from the model-generated classifications were compared to these expert-generated gold standard classifications.

2.2. Scenario 1 – Effects of increasing training dataset size

In Scenario 1, we examined the effect of increasing training dataset size on automated classification results using five sizes of training dataset (2344, 4688, 9376, 18752 and 28128 specimens; Table 1) on a single validation dataset. The training dataset sizes represent 4, 8, 16, 32 and 48 times the size of the validation set size in terms of numbers of specimens (i.e., training to validation ratios of 80:20, 89:11, 94:6, 97:3 and 98:2). The CNN model was then independently applied to each training dataset. Training and validation datasets in Scenario 1 (Table S1) were formed from the pool of identified specimens as follows:

Table 1
Size of the training and validation datasets in Scenarios 1 and 2.

		Training	Validation		
Scenario	Ratio of training: validation specimens	No. specimens	No. sample units	No. specimens per sample unit:median [range]	Seabed area (m^2) per sample unit: median [range]
1	80:20	2344	1	586	780
	89:11	4688			
	94:6	9376			
	97:3	18752			
	98:2	28128			
2	91:9	18752	5	393 [375–410]	778 [768–780]
	82:18	18752	10	406 [371–442]	774 [759–782]
	70:30	18752	20	403 [364–431]	774 [764–782]
	61:39	18752	30	398 [368–437]	772 [754–791]
	54:46	18752	40	402 [349–442]	773 [754–792]
	48:52	18752	50	405 [349–447]	775 [759–792]

(1) First, the validation dataset was selected from the pool. It consisted of 586 specimens in 64 tiles that were the subject of a previous study comparing annotations by three experts (the 'common tiles' in Durden et al., 2016a). These specimens were then excluded from the pool from which the training datasets were formed.
(2) For each training dataset size, ten replicate training datasets were generated by randomly selecting specimens by morphotype from the remaining pool without replacement, to optimally achieve an equal number of specimens per morphotype. This was done to reduce the effects of the class imbalance (Langenkämper et al., 2019). Some imbalance in the training datasets remained, as rarer morphotypes did not achieve the balanced target, consequently the number of morphotypes achieving the balanced target number of specimens declined as the overall training dataset size increased. This process was repeated for each training dataset size, generating 10 replicates for 5 sample sizes.

2.3. Scenario 2 – Effects of increasing validation dataset size

In Scenario 2, we tested the effects of increasing validation set sizes given a fixed training dataset size. The training dataset size (number of specimens = 18752) was selected based on the results of Scenario 1. The training and validation datasets were formed (Table S2) as follows:

(1) Validation datasets were based on replicate sample units from the abyssal plain environment having a sufficient seabed area to generate ecologically appropriate replicates, as determined in Durden et al. (2020b, this Collection; depth group 6), and using the constituent tiles and specimens within each sample unit. The tiles in a typical sample unit represented 775 m^2 seabed area and 405 specimens (Table 1). Validation datasets of six sizes were formed, comprising 5, 10, 20, 30, 40 and 50 sample units. Randomly selected sample units were assigned, without replacement, to the validation datasets, with specimens from the corresponding tiles forming the validation data. Specimens assigned to the validation dataset were removed from the pool of specimens available to form the corresponding training dataset.
(2) For each validation dataset size, the training dataset was formed by randomly selecting from the specimens in the remaining pool without replacement, aiming for the numbers of specimens of each morphotype to approximate the natural community composition (Table S2). This contrasts with Scenario 1, where this class imbalance was minimised.

To compare the impact of validation dataset size, analyses were conducted on a per sample unit basis, using the results from a single CNN run for each validation dataset size. Comparisons were also made to the gold standard expert identifications for the same sample units.

2.4. Comparison of compensation for class imbalance in training data

A comparison of the effects of the two contrasting methods of addressing class imbalance employed in Scenarios 1 and 2 was conducted using validation data generated with the same training dataset size. Results from the validation data (586 specimens) generated with the second-largest training dataset (n = 18752) from Scenario 1 was treated as a single sample unit. This was compared to results from the validation data from 5 randomly-selected sample units in Scenario 2. The corresponding gold standard data were used as the benchmark.

2.5. Specimen preparation

Specimens were extracted from the tiles (Fig. 1). The centre x- and y-pixel coordinates and pixel size of each specimen were used to generate a 'crop' adjusted to the size of the specimen (minimum crop dimensions 75 pixels × 75 pixels). Where pixel coordinates did not correspond to the centre of the specimen, the cropped size was increased to 175% of the measured specimen pixel size. The crops were rescaled to normalise resolution prior to training and validation. Bicubic interpolation was used to resample each crop to a fixed spatial resolution (128 pixels × 128 pixels). To reduce the distortion introduced by resampling, each crop was scaled to a spatial resolution close to the average resolution of the entire dataset (127 pixels × 127 pixels). The exact resolution was selected by rounding the average to the nearest power of 2, to enable more efficient computation. Crops where a portion of the specimen was optically distorted or not visible were removed (e.g. where a specimen was at the edge of an image).

2.6. Details of CNN application

We applied the popular GoogLeNet (Szegedy et al., 2015) model to learn the visual characteristics of the morphotypes in the prepared specimen crops in each training dataset. Implementations of the model that support processing using Tensorflow (Abadi et al., 2016) were used. A pre-trained model was not used, as the input image resolution of existing models is typically much higher (e.g. ImageNet has an input resolution of 299 × 299 pixels; Deng et al., 2009). This would have required the application of large scaling factors to many of the crops, which could have introduced image artefacts affecting the classification process. Data augmentation procedures were applied to the crops in the training datasets to produce additional data by rotating, translating, and filtering (via Gaussian blur) each crop, to aid in producing a robust model while reducing overfitting. The models were trained using a Nvidia Quadro M6000 24 GB graphical processing unit and run for 240 epochs (the number of passes of the training data through the model). The trained model assigned a predicted morphotype (known as 'label' in computer vision nomenclature) to each specimen in each validation set. All programming code was produced using Python 2.7.12 (Rossum, 1995), NumPy 1.14.3 (Oliphant, 2018), and Tensorflow 1.2.1 (Abadi et al., 2016).

2.7. CNN performance evaluation metrics

2.7.1. Computer vision metrics of classification

Classification performance was measured through machine learning metrics: recall, precision, and the F1-score. A specimen was classified correctly if the morphotype classification predicted by the CNN matched the classification given by the gold standard.

$$Recall = t_p/(t_p + f_n) \quad (1)$$

$$Precision = t_p/(t_p + f_p) \quad (2)$$

$$F1\ \ score = 2(Precision \times Recall)/(Precision + Recall) \quad (3)$$

where t_p is the number of true positives (e.g. specimens of morphotype 1 in the gold standard data that the CNN predicted as morphotype 1), f_n is the number of false negatives (e.g. specimens of morphotype 1 predicted as another morphotype), and f_p is the number of false positives (specimens of another morphotype predicted as morphotype 1). Recall, precision and F1 score were computed per morphotype. Recall (Eq. (1)), also known as 'sensitivity', indicates the ability of the model to correctly classify specimens of a particular morphotype, as designated by the gold standard. The F1 score (Eq. (3)) provides an indication of how well a model is able to both correctly classify specimens within a morphotype and also how well it can differentiate between specimens from other classes using the harmonic mean as a single metric. Comparisons of results for different training dataset sizes were assessed in Scenario 1 (that is, 5 training dataset sizes, with 10 replicates each). In Scenario 2, similar comparisons were made between validation datasets sizes (that is, 5 validation dataset sizes, with 5 randomly selected sample units

each). These comparisons were made using Mood's median test and post hoc testing in R using the RVAideMemoire (Hervé, 2020) and rcompanion (Mangiafico, 2020) packages, with p-value adjustment by the Benjamini and Hochberg (1995) method. Boxplots present the median within a box of the interquartile range. Upper and lower whiskers indicate the largest and smallest values no further than 1.5 × interquartile range from each hinge. Outliers (i.e., values beyond whiskers) are presented as points.

2.7.2. Ecological metrics

Hill's indices (N_q, morphotype richness, q = 0; exponential of the Shannon index, q = 1; inverse of Simpson's index, q = 2) (Magurran, 2013), and rarefied morphotype richness (to 300 specimens, EM_{300}) were calculated as measures of diversity (all with units of number of morphotypes; Chao et al., 2014). Comparisons of results for different training and validation dataset sizes were assessed using ANOVA, with post hoc testing using Tukey's honest significant difference method.

Numerical densities of morphotypes were derived from counts and the calculated seabed area of each tile within a sample unit. Differences in apparent community composition between training dataset sizes and the corresponding gold standard were assessed using multivariate statistics (Bray-Curtis dissimilarity measure and 2-dimensional non-metric multidimensional scaling ordination). Numerical densities (individuals ha^{-1}) were subject to several transformations (none, square root, and log[x + 1]) prior to the calculation of dissimilarity measures to assess different aspects of training dataset size variations, giving greater (log[x + 1] transformation) and lesser (no transformation) weight to rare morphotypes; these comparisons were tested using ANOSIM and SIMPER routines (Clarke, 1993).

The precision of univariate diversity measure estimates was quantified as the coefficient of variation. Measurement bias in the diversity measures was assessed as the percentage difference from the corresponding gold standard-derived value. Relative precision in community composition was quantified as multivariate dispersion (Anderson, 2006) among community composition data derived from the CNN results at different training set sizes. Bias and variance in community composition in comparison to the gold standard were computed as the distances between the centroids and the areas of the 2-dimensional non-metric multidimensional scaling ordinations, respectively. All ecological metrics were calculated using the vegan package in R (Oksanen et al., 2012).

3. Results

3.1. Scenario 1: Effects of training dataset size on resulting classified data

3.1.1. Computer vision metrics

The fraction of all specimens that were classified correctly by the CNN ranged from 0.66 to 0.96, and significantly increased with larger training dataset sizes ($\chi^2[4] = 40$, $p < 0.001$). It was at least 0.94 in all model runs for the two largest training dataset sizes. Median recall per training dataset size ranged from 0.67 to 0.95, and was significantly different between training dataset sizes ($\chi^2[4] = 88$, $p < 0.001$; Fig. 2). Post hoc testing showed no significant differences in the largest three training datasets. Median precision and F1-score per training dataset size from 0.44 to 0.98 and from 0.52 to 0.91, respectively, and was significantly different between training dataset sizes ($\chi^2[4] = 158$, $p < 0.001$ and $\chi^2[4] = 185$, $p < 0.001$). Post hoc testing showed no significant differences in precision or F1-score in the largest two training datasets.

The classification success of two common morphotypes, Ophiuroidea and *Iosactis vagabunda* (111 and 308 specimens in the validation dataset, respectively; Table S1), increased as the overall training dataset size was increased, in terms of recall ($\chi^2[4] = 40$, $p < 0.001$ in both cases), precision ($\chi^2[4] = 16$, $p < 0.01$; $\chi^2[4] = 17.6$, $p < 0.01$; respectively), and F1 score ($\chi^2[4] = 40$, $p < 0.001$ in both cases). The recall for these morphotypes reached 0.97 and 0.98, respectively, in the largest two training dataset sizes. The number of training specimens increased as the overall training dataset size increased for these morphotypes (Table S1). Posthoc testing showed no difference between the two largest training dataset sizes for recall in Ophiuroidea, and for F1-score in both Ophiuroidea and *I. vagabunda*.

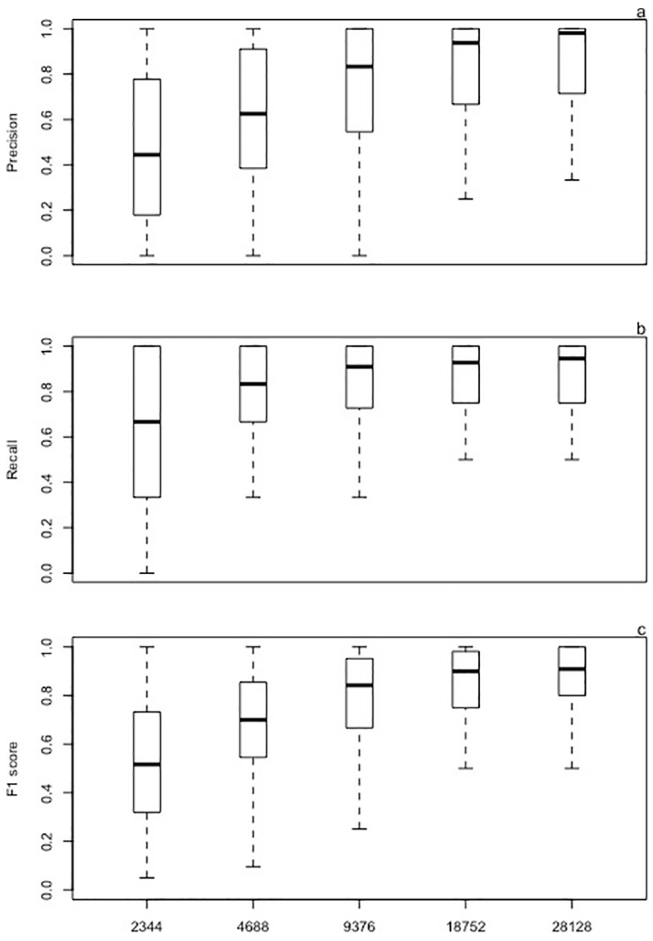

Fig. 2. Scenario 1. Variation in (a) precision, (b) recall and (c) F1-score across all specimens in the CNN-generated classifications by morphotype, with increasing training dataset size (number of training specimens listed).

By contrast, for two uncommon morphotypes, *Porcupinella* sp. and *Psychropotes longicauda* (5 and 4 specimens in the validation dataset, respectively), aspects of classification success improved with increased training dataset sizes, despite the number of training specimens for these morphotypes remaining essentially constant across the training dataset sizes. The classification of *Porcupinella* sp. significantly increased with overall training dataset size (recall $\chi^2[4] = 28$, $p < 0.001$; precision $\chi^2[4] = 24.8$, $p < 0.001$; F1-score $\chi^2[4] = 31.6$, $p < 0.001$), as did the precision of classification of *Psychropotes longicauda* (precision $\chi^2[4] = 20$, $p < 0.001$).

Cnidaria sp.9 provides a case between these extremes, as a moderately common morphotype (21 specimens in the validation dataset), for which the number of training specimens was identical to those of Ophiuroidea and *Iosactis vagabunda* in the smallest three training dataset sizes, but then remained constant at a lesser value for the largest training dataset sizes. The classification of Cnidaria sp.9 was significantly improved by the increase in training dataset size in the latter two training dataset sizes (recall $\chi^2[4] = 36.7$, $p < 0.001$; precision $\chi^2[4] = 40$, $p < 0.001$; F1-score $\chi^2[4] = 40$, $p < 0.001$). Post hoc testing indicated that there was no significant difference in recall or F1-score between the two largest training dataset sizes.

3.1.2. Ecological metrics – diversity

The total number of morphotypes recorded in the data derived from the CNNs ranged from 22 to 24, while the gold standard contained 25 morphotypes. Morphotype richness (N_0) varied significantly ($F[4,45] = 3.0$, $p < 0.05$; Fig. 3) and increased with training dataset size. Hill's diversity numbers N_1 and N_2 varied significantly ($F[4,45] = 54.1$, $p < 0.001$; $F[4,45] = 38.9$, $p < 0.001$, respectively) and decreased with increased training dataset size, but were not significantly different between the three largest training datasets. Rarefied morphotype richness (EM_{300}) was significantly different between training dataset sizes ($F[4,45] = 11.5$, $p < 0.001$).

As training dataset size increased, the coefficients of variation in the Hill's numbers decreased (Fig. 4), and bias in their values from gold standard reduced (Fig. 4). Similarly, the coefficient of variation in rarefied morphotype richness (EM_{300}) decreased with increase in training dataset size, and the bias over the gold standard value became smaller in magnitude, though it was near-identical in the three largest training dataset sizes.

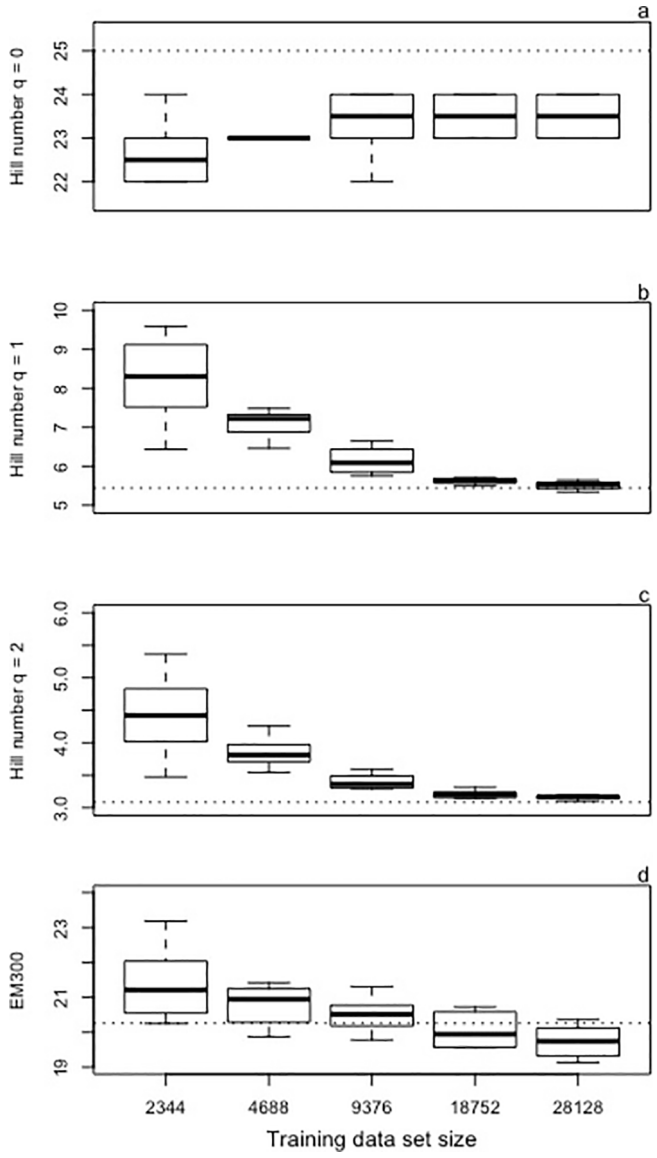

Fig. 3. Scenario 1. Variation in diversity metrics in CNN-generated classifications with training dataset size (number of training specimens listed) by sample unit: (a–c) Hill's indices (q = 0, 1, 2), (d) rarefied morphotype richness (n = 300). Dotted lines represent values from the expert-generated gold standard.

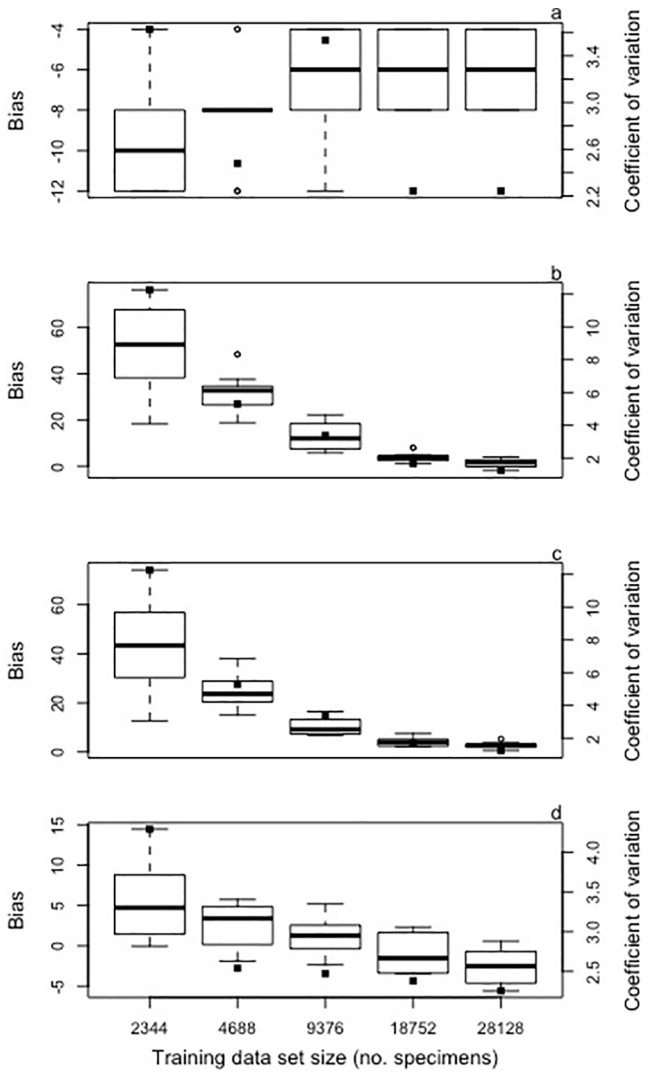

Fig. 4. Scenario 1. Variation in precision (coefficient of variation = CV%; black squares) and bias (mean and range of percentage difference; boxplots) in (a–c) Hill's diversity indices (units are number of morphotypes) and (d) rarefied morphotype richness (EM_{300}) computed per sample unit using CNN-generated classifications with respect to those computed from an expert-generated gold standard with increasing training dataset size (expressed as number of training specimens).

3.1.3. Ecological metrics – faunal composition

Apparent community composition was significantly different among training dataset sizes for all transformations (ANOSIM log[x + 1] R = 0.51, p = 0.001; square root R = 0.52, p = 0.001; no transform R = 0.50, p = 0.001; Fig. 5). Community similarity to the gold standard increased with training dataset size, as reduced bias and variance (Fig. 5d). Differences in multivariate dispersion of community composition between training dataset sizes were significant regardless of the transformation (log[x + 1]-transformed data: ANOVA $F[4,45] = 35.7$, $p < 0.001$; square root-transformed data: $F[4,45] = 35.9$, $p < 0.001$; untransformed data: $F[4,45] = 35.7$, $p < 0.001$). Post hoc testing revealed no significant difference in multivariate dispersion between the largest three training dataset sizes.

Density-driven variations in community composition were detected. Two morphotypes contributed the most to the dissimilarity between successively larger training dataset sizes (untransformed density data): *Iosactis vagabunda* (contributing 12–18%) and Ophiuroidea. These two morphotypes also contributed substantially to the dissimilarity between the untransformed results from the CNNs and the gold standard, with

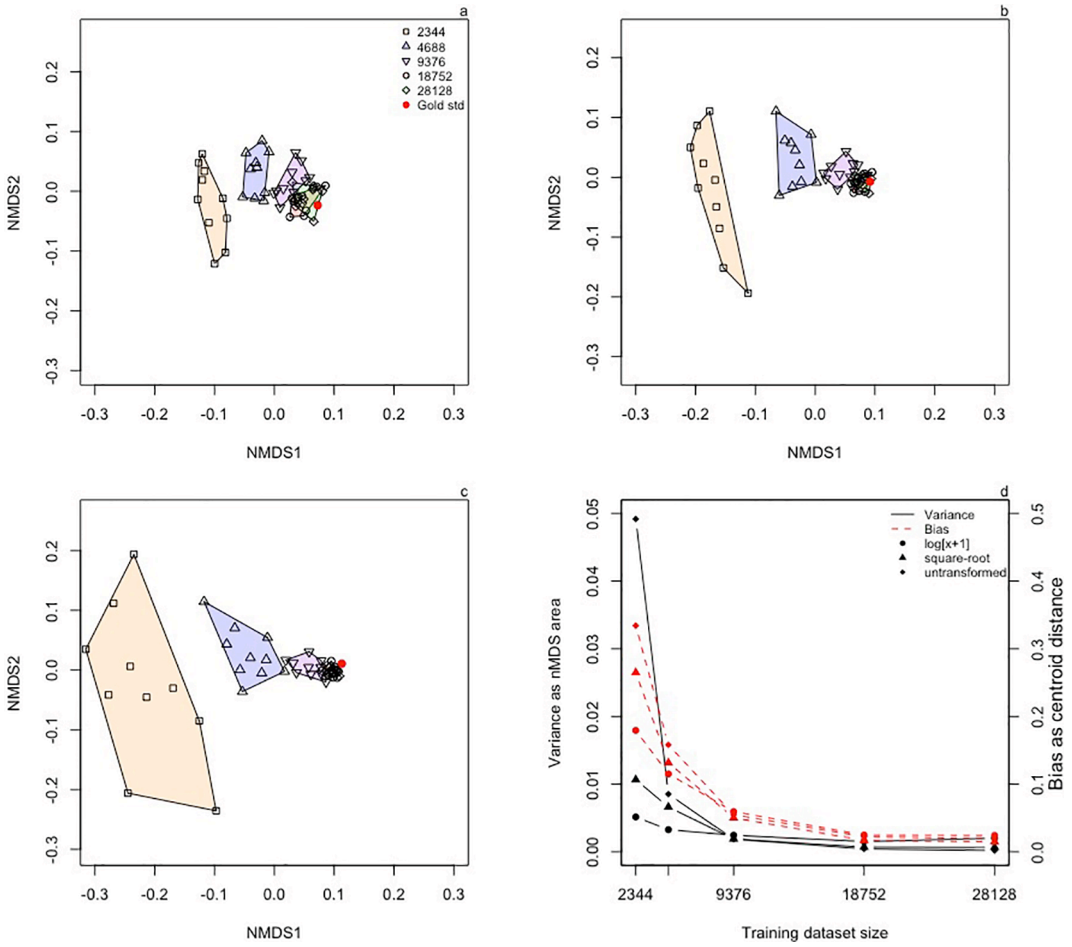

Fig. 5. Scenario 1. Two-dimensional non-metric multidimensional scaling ordination plots of community composition in the CNN-generated classifications by sample unit illustrating variation with training dataset size, and the corresponding expert-generated gold standard: (a) log [x + 1]-transformed, (b) square root-transformed and (c) un-transformed data, and (d) bias (as centroid distance) and variance (as point spread) in community composition from a–c.

I. vagabunda contributing most to the dissimilarity (23–29%) in all training dataset sizes except the largest, where Ophiuroidea contributed most to dissimilarity (22%).

3.2. Scenario 2: Effects of varying validation dataset size

3.2.1. Computer vision metrics

Recall, precision and F1-scores were not significantly different among validation dataset sizes (all $p > 0.05$). Median recall per validation dataset size ranged from 0.93 to 0.99, median precision from 0.99 to 1.0, and median F1-score from 0.92 to 0.94. Recall and precision of Ophiuroidea were significantly different with validation dataset size (both $\chi^2[5] = 12.4$, $p < 0.05$), but did not appear to be systematically related to increasing validation dataset size. None of the computer vision metrics was significantly different with validation dataset size for the remaining morphotypes of interest in Scenario 1: *Iosactis vagabunda*, *Psychropotes longicauda* or *Porcupinella* sp. (all $p > 0.05$).

3.2.2. Ecological metrics – diversity

Rarefied richness (EM_{300}) in the CNN-generated data was not significantly different between the validation dataset sizes (Figure S1), nor from the gold standard. Diversity indices (N_0, N_1, N_2) were not significantly different between validation dataset sizes (all $p > 0.05$), nor different from the gold standard. Neither the coefficients of variation in rarefied richness, the Hill's numbers, nor the bias in these metrics with respect to the gold standard values appeared to vary systematically with validation dataset size (Figure S2).

3.2.3. Ecological metrics – faunal composition

Apparent community composition was not significantly different between validation dataset sizes, regardless of the data transformation applied (log[x + 1], square root or untransformed; ANOSIM $p > 0.05$), nor were statistically significant differences detected from the gold standard (all transformations $p > 0.05$; Fig. 6). Dispersion of community composition was not significantly different between validation dataset sizes (all data transformations $p > 0.05$).

3.3. Effects of class imbalance in training dataset

3.3.1. Computer vision metrics

The overall classification success in the comparison data from Scenario 1 (near-balanced training data) was 0.95, and ranged from 0.94 to 0.97 in Scenario 2 (training specimens proportional to the community composition). In general, recall, precision and F1-scores in Scenario 1 were at least as great as the corresponding values in Scenario 2, with greater discrepancies for the less common morphotypes of interest (*Psychropotes longicauda* and *Porcupinella* sp.; Fig. 7a–c). Differences in the computer vision metrics were slight, with values appearing to be potentially reduced in Scenario 1, for the two common morphotypes (*Iosactis vagabunda* and Ophiuroidea).

3.3.2. Ecological metrics

Bias in morphotype richness over the gold standard was generally negative in Scenarios 1 and 2 (Fig. 7d). Bias in the Hill's numbers (N_1 and N_2) over the gold standard was apparently generally higher and

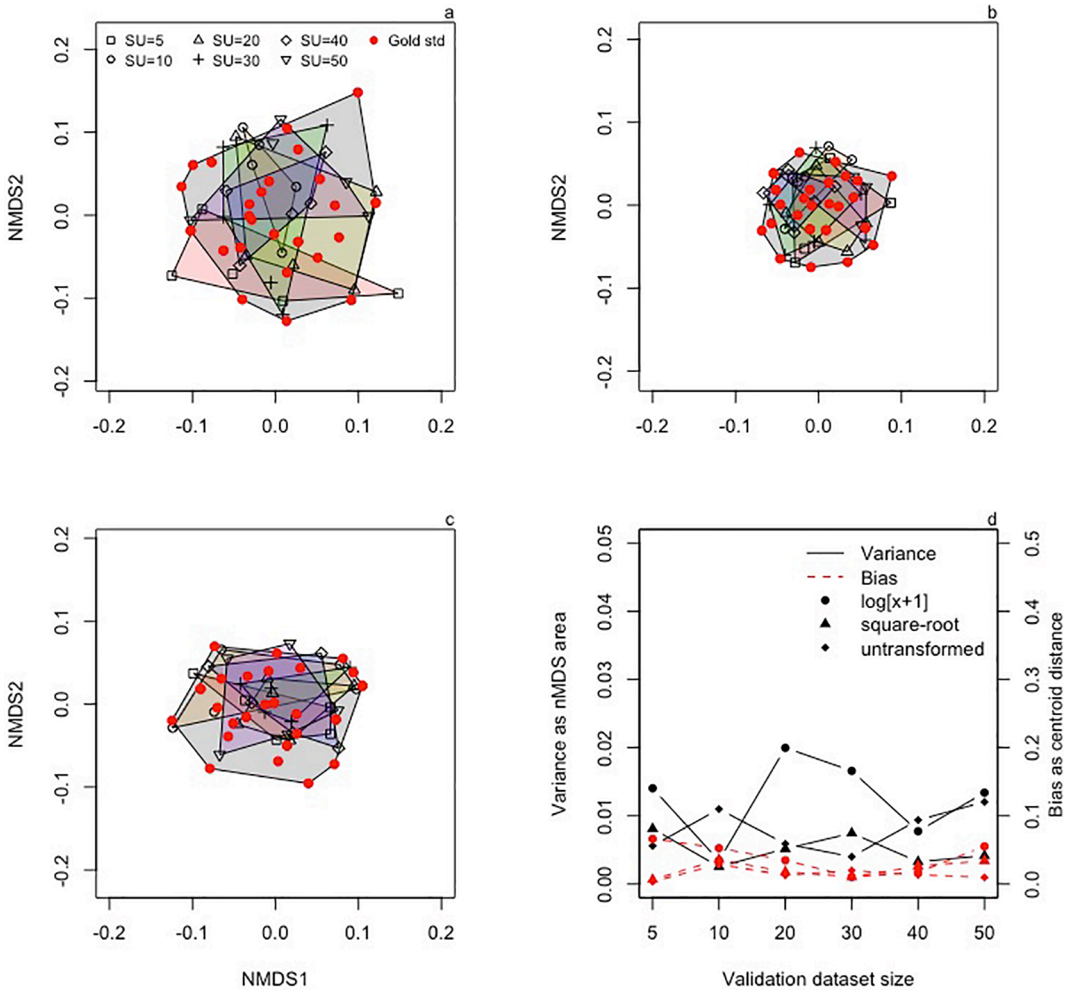

Fig. 6. Scenario 2. Two-dimensional non-metric multidimensional scaling ordination plots of community composition in the CNN-generated classifications by sample unit illustrating variation with validation dataset size, and the corresponding expert-generated gold standard: (a) log[x + 1]-transformed, (b) square root-transformed and (c) un-transformed data, and (d) bias (as centroid distance) and variance (as point spread) in community composition from a–c.

positive in Scenario 1, and either negative or neutral in Scenario 2. That is, the results of Scenario 1 overestimated N_1 and N_2, while Scenario 2 underestimated N_1. Dissimilarity in community composition from the gold standard was very low in Scenario 1, and higher in Scenario 2. It was highest in the log(x + 1) transformed data, where the relative abundances of less common morphotypes were given greater weight.

4. Discussion

Our study of the application of a CNN classifier to a large expert annotated photographic dataset revealed that: (i) a large training dataset size was key to model performance as assessed by both computer vision and ecological metrics, and (ii) the model results were little affected by increasing the validation dataset size. These latter results, which held for validation datasets across an order of magnitude (1965–20250 specimens), give confidence that validation could be successfully achieved with a relatively small dataset, while the model was applied to a much larger dataset without the results being impacted. This is noteworthy given that ecological metrics from expert-generated annotation diverged in the same large photo dataset (Durden et al., 2016a). Note that the minimum validation dataset size needed to facilitate effective evaluation of an AI application with ecological metrics requires sufficient annotated images to produce multiple replicates (Section 2.3), each of which is representative of the sample population (Durden et al., 2016b). The largest validation dataset employed in this study was slightly (8%) larger in size (number of specimens) than the training dataset, suggesting that the model could successfully be applied to datasets at least equivalent in size to an appropriate training dataset; that is, at a 50:50 training to validation or operational dataset ratio, provided the training dataset was sufficiently large.

The appropriate size of the training dataset selected in Scenario 1 (18752 specimens), with which the model achieved 94% correctly classified specimens and precision in ecological metrics, is substantially larger than those tested in other studies. In this training dataset, the morphotypes with the most specimens had almost four times as many specimens as the maximum number tested by Piechaud et al. (2019; 1000 specimens per morphotype), and the total number of training specimens was approximately double that of Piechaud et al. (2019). The training dataset used by Schoening et al. (2012) was similar in size to the smallest of the training datasets used in the present contribution. The training dataset selected in Scenario 1 was even larger than the numbers of specimens annotated in two recent large photographic datasets (Benoist et al., 2019; Simon-Lledó et al., 2019b). This is not to suggest a specific appropriate training size for other datasets, since appropriate size is related to many factors including model type and community structure, but provides some context of scale. In the present study, similar numbers of specimens to the maximum used in Piechaud et al (2019; i.e. $\geq$ 947 per morphotype) were allocated to six morphotypes in three training dataset sizes (n = 9376, 18752, 28128); average recall for these specimens increased from 0.90 to 0.94 with increased training

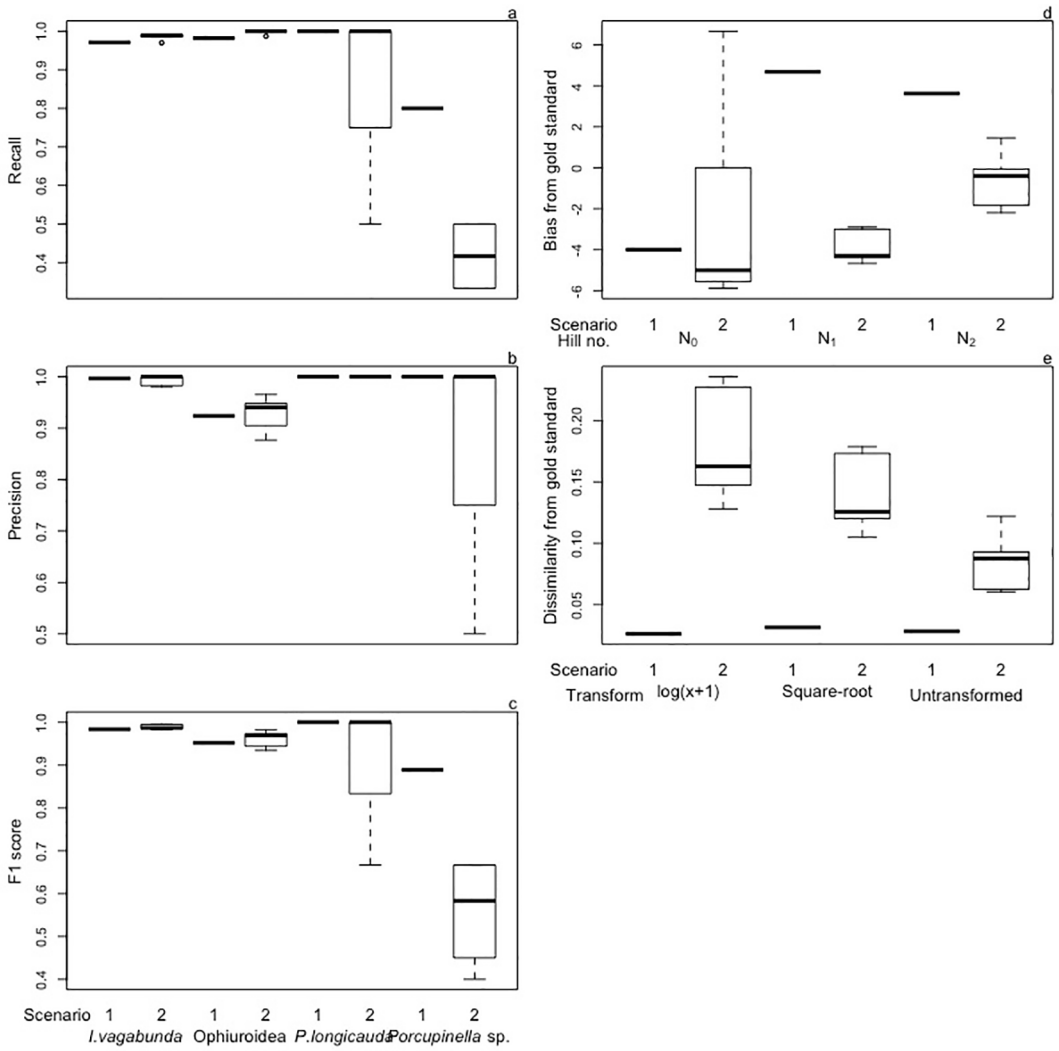

Fig. 7. Comparison of results of differing methods of addressing class imbalance in training dataset, with near-balance in Scenario 1, and proportional to the natural community composition in Scenario 2, with respect to the expert-generated gold standard: (a–c) Recall, precision and F1-score across specimens for four morphotypes of interest (two common: *Iosactis vagabunda* and Ophiuroidea; two uncommon: *Psychropotes longicauda* and *Porcupinella* sp.); (d) bias in diversity metrics per sample unit; and (e) non-metric multidimensional dissimilarity in community composition.

dataset size, significantly higher than the recall achieved in the Piechaud et al. (2019) study. At these training dataset sizes, the resulting diversity metrics had relatively high precision and low bias over the gold standard (Fig. 4), and community composition also converged near the gold standard (Fig. 5). This suggests that much larger training dataset sizes may be required for automated classification of benthic communities than previously anticipated, requiring more human annotation to generate these data prior to AI application.

Increases in the precision of diversity metrics and community composition, and reductions in biases in the same parameters are related to accuracy and recall, as the misclassification of particular morphotypes influences these ecological metrics. From the community composition data in Scenario 1, differences in dispersion between training dataset sizes (Fig. 5) were related to both the identities and numerical densities of the misclassified morphotypes. Recall of the most common morphotypes, *Iosactis vagabunda* and Ophiuroidea, increased 20% and 42% with increased training dataset size, from 0.81 and 0.69 with the smallest training dataset, to 0.97 and 0.98 in the largest training dataset. These two morphotypes contributed most to community dissimilarity between successively larger training dataset sizes, and to differences from the gold standard.

The increase in recall with training dataset size was not limited to morphotypes with increased numbers of training specimens as the overall training dataset size increased. Classification of morphotypes for which the number of training specimens was fewer and remained constant was also improved with increased number of training specimens from other more numerous morphotypes. This benefit to increasing the training dataset size, even for unbalanced communities, suggests that it may be reasonable to include less common morphotypes in automated classification provided that the overall training dataset size is sufficiently large. The inclusion of less common morphotypes can cause increased false positives in the more common morphotypes, reflected in lower precision. However, any elevated false positives in the commonest morphotypes may have been compensated for by the increase in training specimens in the larger training datasets. For example, as the precision of *Iosactis vagabunda* and Ophiuroidea significantly increased with training dataset size, reaching 1.0 and 0.93 with the largest training dataset size. Further improvements to these results may be possible by addressing the class imbalance, for example by applying data augmentation to training specimens non-linearly or through over- or under-sampling (Langenkämper et al., 2019; Qin et al., 2016). More balanced classes in the training data resulted in higher recall, precision and F1-scores for less common morphotypes, minimal bias in diversity metrics over the gold standard, and low dissimilarity from the gold standard.

By contrast, training data that approximated the community composition (a relatively unbalanced training dataset) resulted in reduced recall, precision and F1-scores for less common morphotypes, underestimation of diversity metrics, and greater dissimilarity from the gold standard. This suggests that efforts to reduce class imbalance in the training dataset may improve the quality of the ecological results.

Differences in the resulting ecological metrics between the model-generated data and the gold standard became small in the largest training dataset sizes; however, they remained detectable regardless of validation dataset size. This bias appears to be related to the under recording of morphotype richeness (i.e., N_0 and EM_{300}). While the model-generated results in Scenario 2 were not significantly different from one another regardless of validation dataset size, the modest bias in these results with respect to the gold standard data may be important to consider, in terms of the ecological conclusions and for potential model improvements.

Demand for machine vision and artificial intelligence methods for the assessment of marine ecosystems is likely to grow across a wide range of applications, from spatial and temporal studies at individual locations to extensive, autonomous ocean observing systems (Levin et al., 2019). Effective and cost-efficient monitoring of environmental impacts from industrial activities, such as deep-sea mining (Simon-Lledó et al., 2019a), fishing (Huvenne et al., 2016) and litter (Pham et al., 2014), and corresponding environmental management measures (Benoist et al., 2019; Huvenne et al., 2016) are certain to depend on machine classification in marine imagery from autonomous platforms. Such systems may be employed to automate the classification of the environment (Zelada Leon et al., 2020) and its wildlife (Schneider et al., 2020) to establish baseline conditions or monitor change. Thus, it is timely to develop both machine vision techniques and the methods for their performance assessment in ecosystem applications.

As a developing method, the quantification of bias and precision are important to the interpretation of results generated by it to discern the true biological trend. Continued evaluation of the method will be important as camera technologies and artificial intelligence techniques develop, and particularly in applications to environmental monitoring (Aguzzi et al., 2019), where faunal change over time will be important to detect. Such faunal change may be represented by shifts in the abundances of previously recorded taxa (e.g., Aguzzi et al., 2012; Billett et al., 2010) or involve previously unrecorded taxa, as a result of processes such as biogeographic range shifts or invasions (e.g., Pinsky et al., 2020). It would also impact the selection of training and validation sets, linked to our comments on rare taxa and class imbalance. Thus, the development of artificial intelligence techniques for extraction of ecological data, and the evaluation of such techniques, will require a close collaboration between the computer vision, marine ecology, and environmental policy communities.

5. Conclusions

Our results show that the selection of an appropriate training dataset size is key to ensuring robust CNN-generated classifications of megafauna in seabed photographs, and that an appropriate training dataset size may be very large (e.g., tens of thousands of specimens). We also show that once an appropriate training dataset size is achieved, the model may be applied to an operational dataset of at least equal size to the training dataset, with consistent results. Importantly, these conclusions hold for ecological metrics as well as computer vision metrics. Thus, validation may be conducted on a relatively small dataset, provided that it is large enough to contain multiple appropriately-sized ecological sample units, with confidence that similar results will be obtained in a larger operational dataset. We show that automated classification of less common morphotypes may be feasible, provided that the overall training dataset size is sufficiently large. In addition, we note that tactics for reducing class imbalance in training datasets may result in improvements to the resulting ecological metrics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the captain, crew, and marine autonomous and robotic systems team of RRS *Discovery* cruise 377. Primary data (photographs) used in this study were collected using the NOC autonomous underwater vehicle Autosub6000, and are available from the British Oceanographic Data Centre. We thank Daniel Langenkämper and Tim Nattkemper for valuable discussions on the application of artificial intelligence to the assessment of seabed photographs. This work was supported by the UK Natural Environment Research Council (NERC) through the Autonomous Ecological Surveying of the Abyss project (NE/H021787/1) and the Climate Linked Atlantic Sector Science project (NE/R015953/1). It was also supported by the European Union's Horizon 2020 research and innovation programme projects EMSO-Link (No. 731036), STEMM-CCS (No. 654462) and iAtlantic (No. 818123).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pocean.2021.102612.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv. arXiv: 1603.04467v2.

Aguzzi, J., Chatzievangelou, D., Marini, S., Fanelli, E., Danovaro, R., Flogel, S., Lebris, N., Juanes, F., De Leo, F.C., Del Rio, J., Thomsen, L., Costa, C., Riccobene, G., Tamburini, C., Lefevre, D., Gojak, C., Poulain, P.M., Favali, P., Griffa, A., Purser, A., Cline, D., Edgington, D., Navarro, J., Stefanni, S., D'Hondt, S., Priede, I.G., Rountree, R., Company, J.B., 2019. New High-Tech Flexible Networks for the Monitoring of Deep-Sea Ecosystems. Environ. Sci. Technol. 53, 6616–6631. https://doi.org/10.1021/acs.est.9b00409.

Aguzzi, J., Company, J.B., Costa, C., Matabos, M., Azzurro, E., Màˆnuel, A., Menesatti, P., Sardà, F., Canals, M., Delory, E., Cline, D., Favali, P., Juniper, S.K., Furushima, Y., Fujiwara, Y., Chiesa, J.J., Marotta, L., Bahamon, N., Priede, I.G., 2012. Challenges To The Assessment Of Benthic Populations And Biodiversity As A Result Of Rhythmic Behaviour: Video Solutions From Cabled Observatories. Oceanogr. Mar. Biol. Annu. Rev. 50, 235–286.

Anderson, M.J., 2006. Distance-based tests for homogeneity of multivariate dispersions. Biometrics 62, 245–253. https://doi.org/10.1111/j.1541-0420.2005.00440.x.

Beijbom, O., Edmunds, P.J., Roelfsema, C., Smith, J., Kline, D.I., Neal, B.P., Dunlap, M.J., Moriarty, V., Fan, T.-Y., Tan, C.-J., Chan, S., Treibitz, T., Gamst, A., Mitchell, G., Kriegman, D., 2015. Towards Automated Annotation of Benthic Survey Images: Variability of Human Experts and Operational Modes of Automation. PLoS ONE 10, e0130312. https://doi.org/10.1371/journal.pone.0130312.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Stat. Soc.: Ser. B (Methodol.) 57, 289–300. https://doi.org/10.2307/2346101.

Benoist, N.M.A., Morris, K.J., Bett, B.J., Durden, J.M., Huvenne, V.A.I., Le Bas, T.P., Wynn, R.B., Ware, S.J., Ruhl, H.A., 2019. Monitoring mosaic biotopes in a marine conservation zone by autonomous underwater vehicle. Conserv. Biol. 33, 1174–1186. https://doi.org/10.1111/cobi.13312.

Bett, B.J., 2019. Megafauna. In: Cochran, J.K. (Ed.), Encyclopedia of Ocean Sciences. Elsevier, pp. 735–741.

Bett, B.J., Malzone, M.G., Narayanaswamy, B.E., Wigham, B.D., 2001. Temporal variability in phytodetritus and megabenthic activity at the seabed in the deep Northeast Atlantic. Progr. Oceanogr. 50, 349–368.

Billett, D.S.M., Bett, B.J., Reid, W.D.K., Boorman, B., Priede, I.G., 2010. Long-term change in the abyssal NE Atlantic: The 'Amperima Event' revisited. Deep-Sea Res. Part II: Top. Stud. Oceanogr. 57, 1406–1417. https://doi.org/10.1016/j.dsr2.2009.02.001.

Billett, D.S.M., Bett, B.J., Rice, A.L., Thurston, M.H., Galeron, J., Sibuet, M., Wolff, G.A., 2001. Long-term change in the megabenthos of the Porcupine Abyssal Plain (NE Atlantic). Progr. Oceanogr. 50, 325–348.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., Ellison, A.M., 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monogr. 84, 45–67.
Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18, 117–143.
Danovaro, R., Fanelli, E., Aguzzi, J., Billett, D., Carugati, L., Corinaldesi, C., Dell'Anno, A., Gjerde, K., Jamieson, A.J., Kark, S., McClain, C., Levin, L., Levin, N., Ramirez-Llodra, E., Ruhl, H., Smith, C.R., Snelgrove, P.V.R., Thomsen, L., Van Dover, C.L., Yasuhara, M., 2020. Ecological variables for developing a global deep-ocean monitoring and conservation strategy. Nat. Ecol. Evol. 4, 181–192. https://doi.org/10.1038/s41559-019-1091-z.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, FL, USA, pp. 248–255.
Durden, J.M., Bett, B.J., Huffard, C.L., Pebody, C., Ruhl, H.A., Smith, K.L., 2020a. Response of deep-sea deposit-feeders to detrital inputs: A comparison of two abyssal time-series sites. Deep Sea Res. Part II 173, 104677. https://doi.org/10.1016/j.dsr2.2019.104677.
Durden, J.M., Bett, B.J., Jones, D.O.B., Huvenne, V.A.I., Ruhl, H.A., 2015. Abyssal hills – hidden source of increased habitat heterogeneity, benthic megafaunal biomass and diversity in the deep sea. Progr. Oceanogr. 137 (Part A), 209–218. https://doi.org/10.1016/j.pocean.2015.06.006.
Durden, J.M., Bett, B.J., Ruhl, H.A., 2020b. Subtle variation in abyssal terrain induces significant change in benthic megafaunal abundance, diversity and community structure. Progr. Oceanogr. 186, 102395 https://doi.org/10.1016/j.pocean.2020.102395.
Durden, J.M., Bett, B.J., Schoening, T., Morris, K.J., Nattkemper, T.W., Ruhl, H.A., 2016. Comparison of image annotation data generated by multiple experts for benthic ecology. Mar. Ecol. Prog. Ser. 552, 61–70. https://doi.org/10.3354/meps11775.
Durden, J.M., Ruhl, H.A., Pebody, C., Blackbird, S.J., van Oevelen, D., 2017. Differences in the carbon flows in the benthic food webs of abyssal hills and the plain. Limnol. Oceanogr. 62, 1771–1782.
Durden, J.M., Schoening, T., Althaus, F., Friedman, A., Garcia, R., Glover, A., Greniert, J., Jacobsen Stout, N., Jones, D.O.B., Jordt-Sedlazeck, A., Kaeli, J.W., Koser, K., Kuhnz, L., Lindsay, D., Morris, K.J., Nattkemper, T.W., Osterloff, J., Ruhl, H.A., Singh, H., Tran, M., Bett, B.J., 2016b. Perspectives in visual imaging for marine biology and ecology: from acquisition to understanding. In: R.N. Hughes, D.J. Hughes, I.P. Smith, A.C. Dale (Eds.), Oceanography and Marine Biology: An Annual Review, Vol. 54. CRC Press, pp. 1–72.
Hartman, S., Bett, B.J., Durden, J.M., Henson, S.A., Iversen, M., Jeffreys, R.M., Horton, T., Lampitt, R., Gates, A.R., 2021. Enduring science: three decades of observing the Northeast Atlantic from the Porcupine Abyssal Plain Sustained Observatory (PAP-SO). High resolution temporal and spatial study of the benthic biology and geochemistry of a North-Eastern Atlantic abyssal locality (BENGAL), 191, 102508. 10.1016/j.pocean.2020.102508.
He, K., Gkioxari, G., Dollar, P., Girshick, R., 2020. Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. 42, 386–397. https://doi.org/10.1109/TPAMI.2018.2844175.
Hervé, M., 2020. RVAideMemoire: Testing and Plotting Procedures for Biostatistics.
Hu, Q., Davis, C., 2005. Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine. Mar. Ecol. Prog. Ser. 295, 21–31. https://doi.org/10.3354/meps295021.
Huvenne, V.A.I., Bett, B.J., Masson, D.G., Le Bas, T.P., Wheeler, A.J., 2016. Effectiveness of a deep-sea cold-water coral Marine Protected Area, following eight years of fisheries closure. Biol. Conserv. 200, 60–69. https://doi.org/10.1016/j.biocon.2016.05.030.
Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: 14th International Joint Conference on Artificial Intelligence (IJCAI), Vol. 2. Morgan Kaufmann Publishers Inc., Montreal, Canada, pp. 1137–1143.
Langenkämper, D., Nattkemper, T.W., 2016. COATL – A learning architecture for online real-time detection and classification assistance for environmental data. In: 2016 23rd International Conference on Pattern Recognition. IEEE, Cancun, Mexico, pp. 597–602.
Langenkämper, D., van Kevelaer, R., Nattkemper, T.W., 2019. Strategies for Tackling the Class Imbalance Problem in Marine Image Classification. In: Z. Zhang, D. Suter, Y. Tian, A.B. Albu, N. Sidere, H.J. Escalante (Eds.), Pattern Recognition and Information Forensics: ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA Beijing, China, August 20-24, 2018 Revised Selected Papers, Vol. LNCS 11188. Beijing, China.
Langenkämper, D., van Kevelaer, R., Purser, A., Nattkemper, T.W., 2020. Gear-Induced Concept Drift in Marine Images and Its Effect on Deep Learning Classification. Front. Mar. Sci. 7 https://doi.org/10.3389/fmars.2020.00506.
Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791.
Levin, L.A., Bett, B.J., Gates, A.R., Heimbach, P., Howe, B.M., Janssen, F., McCurdy, A., Ruhl, H.A., Snelgrove, P., Stocks, K.I., Bailey, D., Baumann-Pickering, S., Beaverson, C., Benfield, M.C., Booth, D.J., Carreiro-Silva, M., Colaço, A., Eblé, M.C., Fowler, A.M., Gjerde, K.M., Jones, D.O.B., Katsumata, K., Kelley, D., Le Bris, N., Leonardi, A.P., Lejzerowicz, F., Macreadie, P.I., McLean, D., Meitz, F., Morato, T., Netburn, A., Pawlowski, J., Smith, C.R., Sun, S., Uchida, H., Vardaro, M.F., Venkatesan, R., Weller, R.A., 2019. Global Observing Needs in the Deep Ocean. Front. Mar. Sci. 6, 241. https://doi.org/10.3389/fmars.2019.00241.
MacLeod, N., Benfield, M., Culverhouse, P., 2010. Time to automate identification. Nature 467, 154–155.
Magurran, A.E., 2013. Measuring Biological Diversity. Blackwell Publishing, Oxford, UK.
Mangiafico, S., 2020. rcompanion: Functions to Support Extension Education Program Evaluation.
Marini, S., Fanelli, E., Sbragaglia, V., Azzurro, E., Del Rio Fernandez, J., Aguzzi, J., 2018. Tracking Fish Abundance by Underwater Image Recognition. Sci. Rep. 8, 13748. https://doi.org/10.1038/s41598-018-32089-8.
Matabos, M., Hoeberechts, M., Doya, C., Aguzzi, J., Nephin, J., Reimchen, T.E., Leaver, S., Marx, R.M., Branzan Albu, A., Fier, R., Fernandez-Arcaya, U., Juniper, S. K., 2017. Expert, Crowd, Students or Algorithm: who holds the key to deep-sea imagery 'big data' processing? Methods Ecol. Evol. 8, 996–1004. https://doi.org/10.1111/2041-210X.12746.
Mitchell, E.G., Durden, J.M., Ruhl, H.A., 2020. First network analysis of interspecific associations of abyssal benthic megafauna reveals potential vulnerability of abyssal hill community. Progr. Oceanogr. 187, 102401 https://doi.org/10.1016/j.pocean.2020.102401.
Morris, K.J., Bett, B.J., Durden, J.M., Benoist, N.M.A., Huvenne, V.A.I., Jones, D.O.B., Robert, K., Ichino, M.C., Wolff, G.A., Ruhl, H.A., 2016. Landscape-scale spatial heterogeneity in phytodetrital cover and megafauna biomass in the abyss links to modest topographic variation. Sci. Rep. 6, 34080. https://doi.org/10.1038/srep34080.
Morris, K.J., Bett, B.J., Durden, J.M., Huvenne, V.A.I., Milligan, R., Jones, D.O.B., McPhail, S., Robert, K., Bailey, D., Ruhl, H.A., 2014. A new method for ecological surveying of the abyss using autonomous underwater vehicle photography. Limnol. Oceanogr. Methods 12, 795–809. https://doi.org/10.4319/lom.2014.12.795.
Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2012. vegan: Community Ecology Package.
Oliphant, T.E., 2018. NumPy 1.14.3.
Osterloff, J., Nilssen, I., Nattkemper, T.W., 2016. A computer vision approach for monitoring the spatial and temporal shrimp distribution at the LoVe observatory. Methods Oceanogr. 15–16, 114–128. https://doi.org/10.1016/j.mio.2016.03.002.
Pham, C.K., Ramirez-Llodra, E., Alt, C.H., Amaro, T., Bergmann, M., Canals, M., Davies, J., Duineveld, G., Galgani, F., Howell, K.L., Huvenne, V.A.I., Isidro, E., Jones, D.O.B., Lastras, G., Morato, T., Gomes-Pereira, J.N., Purser, A., Stewart, H., Tojeira, I., Tubau, X., Van Rooij, D., Tyler, P.A., 2014. Marine litter distribution and density in European Seas, from the shelves to deep basins. PLoS ONE 9, e95839.
Piechaud, N., Hunt, C., Culverhouse, P.F., Foster, N.L., Howell, K.L., 2019. Automated identification of benthic epifauna with computer vision. Mar. Ecol. Prog. Ser. 615, 15–30. https://doi.org/10.3354/meps12925.
Pinsky, M.L., Selden, R.L., Kitchel, Z.J., 2020. Climate-Driven Shifts in Marine Species Ranges: Scaling from Organisms to Communities. Ann. Rev. Mar. Sci. 12, 153–179. https://doi.org/10.1146/annurev-marine-010419-010916.
Purser, A., Bergmann, M., Lundalv, T., Ontrup, J., Nattkemper, T.W., 2009. Use of machine-learning algorithms for the automated detection of cold-water coral habitats: a pilot study. Mar. Ecol. Prog. Ser. 397, 241–251. https://doi.org/10.3354/meps08154.
Qin, H., Li, X., Liang, J., Peng, Y., Zhang, C., 2016. DeepFish: Accurate underwater live fish recognition with a deep architecture. Neurocomputing 187, 49–58. https://doi.org/10.1016/j.neucom.2015.10.122.
Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.
Rossum, G.V., 1995. Python tutorial. Vol. Technical Report CS-R9526. Amsterdam: Centrum voor Wiskunde en Informatica (CWI), Computer Science/Department of Algorithmics and Architecture.
Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M., Seager, J., Harvey, E., 2016. Fish species classification in unconstrained underwater environments based on deep learning. Limnol. Oceanogr. Methods 14, 570–585. https://doi.org/10.1002/lom3.10113.
Schneider, S., Greenberg, S., Taylor, G.W., Kremer, S.C., 2020. Three critical factors affecting automated image species recognition performance for camera traps. Ecol. Evol. 10, 3503–3517. https://doi.org/10.1002/ece3.6147.
Schoening, T., Bergmann, M., Ontrup, J., Taylor, J., Dannheim, J., Gutt, J., Purser, A., Nattkemper, T.W., 2012. Semi-Automated Image Analysis for the Assessment of Megafaunal Densities at the Arctic Deep-Sea Observatory HAUSGARTEN. PLoS ONE 7, e38179. https://doi.org/10.1371/journal.pone.0038179.
Shafait, F., Mian, A., Shortis, M., Ghanem, B., Culverhouse, P.F., Edgington, D., Cline, D., Ravanbakhsh, M., Seager, J., Harvey, E.S., 2016. Fish identification from videos captured in uncontrolled underwater environments. ICES J. Mar. Sci. 73, 2737–2746. https://doi.org/10.1093/icesjms/fsw106.
Simon-Lledó, E., Bett, B.J., Huvenne, V.A.I., Koser, K., Schoening, T., Greinert, J., Jones, D.O.B., 2019a. Biological effects 26 years after simulated deep-sea mining. Sci. Rep. 9, 8040. https://doi.org/10.1038/s41598-019-44492-w.
Simon-Lledó, E., Bett, B.J., Huvenne, V.A.I., Schoening, T., Benoist, N.M.A., Jeffreys, R. M., Durden, J.M., Jones, D.O.B., 2019b. Megafaunal variation in the abyssal landscape of the Clarion Clipperton Zone. Progr. Oceanogr. 170, 119–133. https://doi.org/10.1016/j.pocean.2018.11.003.
Szegedy, C., Liu, W., Jia, Y., Semanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Boston, MA, USA.
Zelada Leon, A., Huvenne, V.A.I., Benoist, N.M.A., Ferguson, M., Bett, B.J., Wynn, R.B., 2020. Assessing the Repeatability of Automated Seafloor Classification Algorithms, with Application in Marine Protected Area Monitoring. Remote Sens. 12 https://doi.org/10.3390/rs12101572.