2019-04-18

Automated identification of benthic epifauna with computer vision

Piechaud, Nils

http://hdl.handle.net/10026.1/13732

10.3354/meps12925
Marine Ecology Progress Series
Inter Research

Automated identification of benthic epifauna with computer vision

Running head: Applying computer vision to benthic ecology

Authors: Nils Piechaud[1], Christopher Hunt[2], Phil F. Culverhouse[3], Nicola L. Foster[1], Kerry L. Howell[1]

1: School of Biological and Marine Sciences, University of Plymouth, Plymouth, PL4 8AA, UK

2: Controlled frenzy LTD, THINQTANQ, Fairbairn House, Higher Lane, Plymouth, PL1 2AN

3: School of Computing, Electronics and Mathematics, University of Plymouth, Plymouth, PL4 8AA, UK

Corresponding author email: Nils.Piechaud@plymouth.ac.uk

Abstract: Benthic ecosystems are chronically undersampled, particularly in environments >50m. Yet, a rising level of anthropogenic threats makes data collection ever more urgent. Currently, modern underwater sampling tools, particularly Autonomous Underwater Vehicles (AUV), are able to collect vast image datasets, but cannot bypass the bottleneck formed by manual image annotation. Computer Vision (CV) offers a faster, more consistent, cost effective and a sharable alternative to manual annotation. We used Tensorflow to evaluate the performance of the Inception V3 model with different numbers of training images, as well as assessing how many different classes (taxa) it could distinguish. Classifiers (models) were trained with increasing amounts of data (20 to 1000 images of each taxa) and increasing numbers of taxa (7 to 52). Maximum performance (0.78 Sensitivity, 0.75 precision) was achieved using the maximum number of training images but little was gained in performance beyond 200 training images. Performance was also highest with the least classes in training. None of the classifiers had average performances high enough to be a suitable alternative to manual annotation. However, some classifiers performed well for individual taxa (0.95 sensitivity 0.94 precision). Our results suggest this technology is currently best applied to specific taxa that can be reliably identified and where 200 training images offers a good compromise between performance and annotation effort. This demonstrates that CV could be routinely employed as a tool to study benthic ecology by non-specialists, which could lead to a major increase in data availability for conservation research and biodiversity management.

Key words: Benthic Ecology, Computer Vision, Automated Image Analysis, Automated species identification

Introduction

Marine ecosystems cover the majority of Earth's surface but benthic ecologists and biodiversity mangers have long been confronted with a shortage of data (Jongman 2013, Borja et al. 2016) regarding its composition and functioning. With increasing anthropogenic pressure, management measures need to be implemented urgently (Van Dover et al. 2014, Danovaro et al. 2017). These conservation measures must be based on a solid understanding of taxonomic diversity and ecological dynamics of habitats considered (Hernandez et al. 2006). In many cases, that knowledge is lacking and specialists agree that data collection must be increased to tackle the challenge (Costello et al. 2010, Borja et al. 2016). The amount of data currently available on benthic ecosystems is always limited by how many samples can be collected, stored, and processed at a time. Since the 19th century, various technological innovations have attempted to bypass this bottleneck.

Benthic ecosystems are traditionally sampled by trawls, cores and other physical means. These physical samples are costly to collect and process, and logistically challenging to store (Clark et al. 2016). While physical samples remain the mainstay of benthic surveys, use of underwater imaging technologies is increasingly popular among marine ecologists (Solan et al. 2003, Bicknell et al. 2016, Brandt et al. 2016, Romero-Ramirez et al. 2016). These technologies offer a less invasive, more cost effective method of survey, and storage space for image data is virtually unlimited (Mallet & Pelletier 2014). Underwater imaging is now regularly utilised alongside other sampling tools to provide a comprehensive view of the marine environment.

Modern underwater sampling vehicles, and particularly Autonomous Underwater Vehicles (AUV), have great potential in providing the step-change in the rate of data gathering that is needed to support sustainable marine environmental management. They are capable of collecting large numbers of images of the sea bed in a single deployment (Lucieer & Forrest 2016, Williams et al. 2016). For example, a 22 hour AUV dive can deliver more than 150,000 images of the seafloor along with other types of environmental data (Wynn et al. 2012). Comparatively, trawls and Remotely Operated Vehicles (ROV) cover less ground per dive and the ship and its crew are unable to operate any other benthic equipment while they are deployed (Brandt et al. 2016, Clark et al. 2016).

To translate the information contained in images into semantic data that can then be used in statistical analysis, a step of manual analysis (or annotation) is conducted by trained scientists. Human observers, even highly-trained, do not achieve 100%

correct classification rates and are highly inconsistent across time and across annotators (Culverhouse et al. 2003, Culverhouse et al. 2014, Beijbom et al. 2015, Durden et al. 2016). Besides, manual image annotation results are subject to observer bias, meaning interpretations vary depending on the annotators experience and their mood changes across the analysis process (tiredness, boredom or stress, etc…) (Culverhouse et al. 2003, Durden et al. 2016). The results (format, taxonomic resolution and nomenclature) of these analyses also tend to differ from one institution, project or individual annotator to another. This lack of standardisation makes merging and comparing datasets difficult (Bullimore et al. 2013, Althaus et al. 2015, McClain & Rex 2015), and the data quality is not always consistent. More importantly, manual analysis is a time consuming process, which forms the current bottleneck in image based marine ecological sampling (Edgington et al. 2006, Beijbom et al. 2015, Schoening et al. 2017). The growing trend towards use of AUVs for seafloor biological survey will only worsen this situation.

Artificial intelligence (AI) and computer vision (CV) provide potential means by which to both accelerate and standardise the interpretation of image data (Culverhouse et al. 2003, MacLeod et al. 2010, Beijbom et al. 2012, Favret & Sieracki 2016). Although using AI for biological research has a long history (Rohlf & Sokal 1967, Jeffries et al. 1984, Gaston & O'Neill 2004), it has always been challenging to implement for non-specialists and requires skills and materials that most biologists do not have access to (Gaston & O'Neill 2004, Rampasek & Goldenberg 2016).

CV has been successfully applied to benthic species identification by a growing number of studies (Edgington et al. 2006, Beijbom et al. 2015, Marburg & Bigham 2016, Manderson et al. 2017, Norouzzadeh et al. 2018, Schneider et al. 2018) but has yet to be made into an easy to use tool that any biologist in the field can implement as an alternative to manual image annotation and integrate with previous analysis. Multiple potential commercial applications, the availability of new tools as open software, as well as the improvement of hardware capacity are driving new developments in AI (e.g. neural networks and deep learning). This is likely to change how AI can be employed in the field of scientific research (Rampasek & Goldenberg 2016, Weinstein 2018). In parallel, new image analysis and data science software allow an easier and more efficient integration of various tools into the research process, from data collection to final scientific or public outreach material (Gomes-Pereira et al. 2016). These new technologies are potentially enabling full automation of the annotation process and could revolutionise ecological research (Weinstein 2018).

While the principle of automated classification (automated assignation of pre-established classes to objects on images) has been validated, few practical examples exist of AI-based methods used to identify benthic animals from images acquired by AUV. Consequently, implementing an automated species classifier is a potentially time consuming investment for an uncertain return. Relying on proven manual methods remains the safe option for researchers. Practical guidance is needed to help ecologists decide whether adopting AI and CV is feasible and would fit their dataset and scientific objectives.

To make that decision, benthic ecologists need to know:

- What level of accuracy and uncertainty can be expected from CV annotation and does it match or approximate the accuracy of human annotators.
- How much material is needed to train a classifier and is a limited amount obtained from a single study sufficient.

- How to assess their own dataset to decide whether use of CV is appropriate.

In this study, we investigate these issues by using an open access algorithm to build a Convolutional Neural Network (CNN) to identify benthic animals in seafloor images, obtained from a single deployment of the UK's Autosub6000 AUV. Technically speaking, we seek to train an automated classifier that is able to determine which taxa an animal on an image most likely belongs to, using a list of pre-defined taxa (or classes). Specifically we ask, 1) what impact does the number of images, on which the classifier is trained, have on its performance? and 2) What impact does the number of classes, on which the classifier is trained, have on its performance? In addition, we provide a case study in the application of CV to an unbalanced ecological dataset.

Method

Study area and data collection:

All the images used in this study were collected by the UK's national AUV Autosub6000 in May 2016 as part of the NERC funded DeepLinks (JC136) research cruise. The images were taken as part of a 1880 m long transect at station 26 of that cruise at 1200 meters depth on the north-east side of Rockall Bank, N.E. Atlantic. This region was selected for the study due to the flat topography and low likelihood of disturbance, making it ideal for AUV deployment. The AUV was equipped with a downward facing Grasshopper2 GS2-GE-50S5C camera from Point Grey Research. The AUV was flown at 1.1ms^{-1} speed, at 3m ±0.1 m off bottom and took images every second, resulting in near overlapping image coverage. The surface area of each image is between 1 and 2.5 m^{2}, and the resolution is 2448 x 2048 at 5 mega pixels.

In total, 1165 raw photos of the seabed were manually annotated by a single observer with the Biigle 2.0 software (Langenkämper et al. 2017) using a regional catalogue of Operational Taxonomical Units (OTU) developed (Howell & Davies 2016). Within the Biigle 2.0 software, location (X and Y coordinates in pixels within the photo for point annotations, or X, Y and radius for individuals marked using a circle) and identity of individual OTUs annotated within each image was recorded and stored. For each OTU, all individual annotations were visually inspected using the "Largo" evaluation tool in Biigle 2.0, to ensure consistency in identification and reduce error.

Image data

Manual image annotation resulted in a dataset consisting of 41208 individuals belonging to 148 OTUs. Each individual was then cropped from the raw image, together with its assigned OTU label, using a custom Python (www.Python.org) script. For each annotation, a square of 40 pixels or more, positioned manually on X and Y coordinates of the centre of the animal, was fitted and cropped out. For animals bigger than 40 pixels, the size of the square was manually set to encompass the whole individual. These cropped image slices and associated OTU labels (to become classes in the model training design) formed the input used in the CNN.

Tensorflow and transfer learning

Rather than train our own neural network, we used transfer learning (Pan & Yang 2010) to retrain the Inception V3 model (Szegedy et al. 2016), a CNN built in the freely available library Tensorflow (Abadi et al. 2016).

CNNs are a particular architecture of neural networks, more specifically, deep learning, particularly suited to image analysis (Krizhevsky et al. 2012, LeCun et al.

2015). A CNN has the capacity to detect and match patterns in images thereby "learning" what features are relevant to differentiate objects and, subsequently, classify them accordingly.

Tensorflow (TF) is a C++ based library but has a Python Application Programming Interface (API) that makes it easier to train, tune and deploy neural networks. Transfer learning is a method allowing a CNN built on a large dataset to be re-repurposed into a classifier capable of distinguishing between classes it was not initially trained on. The strength of this method is that the dataset on which it is transferred does not need to be as large as it should be to train a CNN from the beginning. Here, we were able to train a classifier with a tens to hundreds of images per class (in our case, OTUs) instead of millions.

Classifier training and testing

A random 75-25% split was applied to every OTU in order to separate images used for training the classifier and those used for testing. The training and test data sets for all OTUs were then combined into single 'training' and 'test' datasets.

The OTUs the classifier was trained to identify are referred to as classes and only those OTUs for which there were a sufficient number of image slices (individual observations) available were selected for use in training. The minimum number of images needed for training was set to 20. This means that for an OTU to be included in the study at least 27 image slices were needed, 20 for training and 7 for testing. Out of the 148 OTUs observed, 52 were above that threshold. The remaining 96 OTUs represented 3.19% of the total number of individual annotations and were removed from the dataset.

The classifier was trained on the training dataset and then predictions were made on the test dataset. For each cropped image slice in the test dataset, TF gave a score for each of the possible OTU classes for which it had been trained. The scores range from 0 to 1 (the sum of scores for all classes being 1) and represent the model's confidence that the slice belongs to the corresponding class. The final prediction was the OTU class that received the highest score. The prediction was then compared to the manually assigned OTU class.

To measure the effect of the number of training images (or limit) on the accuracy and confidence of the predictions, the training data set was filtered so each OTU class was represented by 20, 50, 100, 200, 500, and 1000 images (Table 1). A classifier was then trained on each of these six pools of images and tested using the test data set. Only seven OTUs were frequent enough to be used with these six limits (Figure 1).

The combination of groups and limits is referred to as treatments and designation of each treatment follows the nomenclature in table 1 (e.g. A1000 is group A, limit 1000). Each treatment was repeated 10 times with different random splits between testing and training data.

To measure the effect of the number of OTU classes used to train the CNN on its capacity to correctly classify the test dataset, we used three training datasets each with different numbers of classes (referred to as groups) (Table 1). The number of classes is defined by the number of available images per OTU so classifiers can be trained on a set number of images for every class while retaining enough images for testing. Group A contained 7 classes for which more than 1000 images was available; group B contained 27 classes for which more than a 100 images were available; and group C contained 52 classes for which more than 20 images were available.

Within each group, classifiers were trained with all six pools of images (Table 1).

Note that when the limit is above the available number of images, the classes with less images were trained with the maximum number available regardless of the limit. This results in class imbalance in the model training for some treatments in group C with more than 20 images and in B with more than 100 images (balanced treatments are listed in Table 1). To assess the effect of the number of OTU classes used to train the CNN on its capacity to correctly classify the test dataset only balanced designs were used.

In total, 180 (3x6x10) classifiers were trained and tested. All the CNNs were trained in the Google Cloud ML (https://cloud.google.com/) remote computing facility.

To be applied to a "real life" ecological study, the classifiers have to maximize performances while minimizing the initial effort needed to build the training dataset. To assess appropriate use of CV on a 'real life' dataset we considered all possible combinations of numbers of training image and numbers of OTU classes in an unbalanced design. Average performances and individual OTU performances were assessed.

Analysis and performances evaluation

Considering each class, the observation can be a presence (the OTU is present on the image) or an absence (the OTU is not on the image and another OTU is). The different possible outcomes or predictions of the classifier are detailed in Table 2.

The respective number of each outcome type (the confusion matrix) was used to calculate performance metrics.

The classification accuracy is the percentage of predictions that are correct (prediction matches observation) and is often used to evaluate performances in ML studies. This measure ignores the differences between classes, thus we used two model evaluation metrics which rely on a confusion matrix (Manel et al. 2001) explained in Table 2.

- **Sensitivity,** also referred to as *true positives rate* or *recall*. It varies between 0 and 1. It quantifies the proportion of individuals of a given OTU in the testing set that are correctly identified. A value of 1 means that all individuals of a given OTU are identified as such.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- **Precision,** *or Positive Predictive Value*. It varies between 0 and 1. It quantifies the proportion of true positives among the individual identified as a given OTU. A value of 1 means all the individual identified as a given OTU class are indeed that OTU.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Average and standard deviation for all metrics were calculated for each class within each treatment and then averaged over other grouping factors. This gave an estimation of the overall performance of the classifiers. The performances of the classifiers for each individual class were also carefully analysed.

Differences in metrics were statistically tested with a permutation-based analysis of variance in the “lmPerm” package in R (Wheeler & Torchiano 2010). We report p-values classified with five levels of significance: more than 0.05 or non-significant, less than 0.05, less than 0.01, less than 0.001 and less than 0.0001. Relationships between number of images and performance were extrapolated with a neural network regression in the “nnet” package in R (Ripley et al. 2016) projected over 1000 to 10000 images. All data analyses were carried out in R (Team 2014) using the “tidyverse” package (Wickham 2017).

Results

The results are presented in three sections. First, questions related to the impact of the number of training images are addressed, then the effect of the number of classes in the training set is assessed, and finally the results relevant to choosing the best method in our case study are presented.

Impact of the number of training images on performance

Average performance, measured as both sensitivity and precision, increases with an increasing number of images used (Figure 2). For sensitivity, there is an average increase from 0.64 to 0.78 when moving from 20 to 1000 images, respectively. This is mirrored by increases in precision from 0.63 to 0.75 when moving from 20 to 1000 images, respectively. Non-linear extrapolations of average sensitivity and precision show that performances reached with 1000 training images may be close to an asymptote and performances obtained with additional training material probably plateau below 0.78 for sensitivity and 0.75 for precision (Figure 2). This suggests that the model is unable to achieve perfect performance regardless of how many additional images are used in training.

The number of images has a clear positive effect on performances. For almost all pairs of models compared in (Figure A1), performance values are statistically significantly different ($p < 0.05$) and very often, significance is very high (p-value < 0.0001). There are a few exceptions like between A20 and A50 classifiers where p-value > 0.05 for sensitivity and between 0.01-0.05 for precision or the B1000 classifier, for which there is no significant difference between B500 and B200 in sensitivity. However, measured difference in performance between sequential models becomes vanishingly small at higher numbers of training images, such that the difference between A200 and A1000 classifiers is 0.04 for sensitivity and 0.05 for precision. This suggests little to no improvement is gained in model performance by using more than 200 training images.

There are strong between-OTU differences in classifier performances (Figure 3). All classifiers have high sensitivity for OTU261 and OTU339, even the A20 classifier (0.88 and 0.77, respectively). For OTU2 and OTU23, classifiers have more variable and lower sensitivity regardless of the number of training images used. OTU261 is very constant in shape and colour and has a distinctive pattern on its outside. OTU339 can be in different pose or orientation within an image but has a number of distinguishing features such as its reflective eyes, and its long, often spread-out limbs. OTU2 and OTU23 are both anemones. OTU2 is a cerianthid (a tube anemone) of various size and orientation and OTU23 is a Halcampidae/Edwardsiidea like anemone of very small size.

The OTUs for which precision is highest are not necessarily those for which sensitivity is highest. The highest precision observed was for OTU261 but the second highest precision observed was for OTU603, which has a lower sensitivity. For some classes (OTU261 or OTU339), precision is lower with 50 training images compared to 20 training images.

Impact of the number of classes on classifier performance

Classifiers trained with 7 classes (group A) had significantly better sensitivity (Figure A1) and precision than equivalent classifiers trained on more classes but the same number of images (Figure 4). Variability in performance was also lower for classifiers trained with fewer classes. Average sensitivity decreased from 0.71 to 0.38, and average precision decreased from 0.69 to 0.32, when moving from 7 to 27 classes. This suggests a negative effect of the number of classes on performance; however, on average, there is only a minor drop in performance (0.018 in sensitivity and 0.035 in precision) between classifiers trained on 27 and 52 classes. Interestingly, B100 and C100 both have sensitivity of 0.38 (no statistical difference) and C20 has higher (+ 0.02) sensitivity than B20.

OTUs that perform well in a group tend to perform well in others. OTU261 and OTU339 are in the top 10 for each group although their performances are lower in group B and C.

Application of CV to an unbalanced ecological dataset

When considering all treatments in an unbalanced design (Figure 4), the average sensitivity per treatment ranges from 0.32 to 0.78. The highest sensitivity was achieved by the A1000 classifier (7 classes, with 1000 training images in each class) while the lowest was achieved by the B20 and C20 classifiers (27 and 52 classes, respectively, and 20 images in each class). A1000 also had the highest precision (0.75), with the lowest precision observed in the C20 classifier (0.20). Sensitivity of the C1000 classifier (where class imbalance is highest) was lower than in the C100 and C200 classifiers but precision simply increases with the number of training images, although this could be an artefact driven by the improvement of precision on the most abundant classes.

When considering individual OTUs, performance was unacceptably low for most, but not all as some had sensitivity and precision greater than 0.85. Based on average sensitivity across all treatments, the top 10 and the bottom 10 OTU classes were identified. The top 10 classes were large animals with consistent or distinctive shape, colour and patterning. They were not necessarily the most abundant classes as six of them were only present in group C, for which there are less than 100 training images, and only two in A, for which there are at least 1000 training images. The two of these OTU present in group A had better average precision than any other OTU class in the top 10. Those OTU classes with the worst performances are generally those for which there are fewer training images (group C). They also tend to be smaller organisms, have colours similar to the background and have very variable shapes and sizes.

In this dataset, CV could be applied to OTU261 and OTU339. These OTUs were both very abundant in the study area, justifying automated annotation, and they both had very high performances, making their identification by the classifier reliable (Figure 5).

The performance of CV for OTU261 and OTU339 was maximised in the A1000 classifier with only 7 classes and 1000 training images. The A200 classifier also achieved performances close to A1000 despite being trained on five times less images. For OTU261, even the A20 and A50 classifiers achieved sensitivity and precision greater than 0.86, and differences between the A20, A50 and A100 classifiers were not statistically significant (Figure 5).

Sensitivity in the C1000 classifier was 0.92 and 0.89 for OTU261 and OTU339, respectively, which is significantly lower than A1000 (p-value <0.0001 for both – Figure A2 and A3) but only a marginal difference (0.03 each). For OTU261, the C200 classifier achieved lower sensitivity than the A200 but they had equal precision. For OTU339, precision is also the same in A200 and all C classifiers (Figure A4). Note that for both OTUs, precision of all treatments in C were either not significantly or barely significantly different (p-value above 0.01). Thus, C classifiers (with 52 classes) achieve performances almost as good as A classifiers when training on 200 or less images.

Group B classifiers tended to show slightly lower sensitivity than A classifiers and slightly lower precision than C, although often not significantly different.

Discussion

In this study, our purpose was to test the capacity of a transferred CNN classifier (partially trained on a different dataset) to identify benthic animals and, by extension, to test if this methodology can be successfully applied in ecology by non-specialists with a relatively small data set, open-source software and libraries, as well as a short investment in time after manual image annotation.

Overall performances

Our classifiers achieved maximum average performance of 0.78 in sensitivity and 0.75 in precision. In other studies, performances achieved through manual annotation range from 50 to 95% for benthic fauna (Beijbom et al. 2015, Durden et al. 2016) and 84 to 94% accuracy for plankton (Culverhouse et al. 2003). There is no consensus on what is an acceptable error rate in the ecological literature but, to be competitive with experts, automated identification performances should be towards the higher end of those achieved manually. In this regard, Culverhouse et al. (2014) report an anecdotal value of 0.9 cited by experts. Previous studies on marine ecosystems sampled via images that have attempted to automatically classify multiple benthic megafaunal taxa with various methods sometimes achieve performances comparable to those of experts. Beijbom et al. (2012) found average accuracies up to 97% when classifying different coral species in shallow reefs. Schoening et al. (2012) found an average sensitivity of 0.87 and precision of 67% when classifying deep benthic megafauna in the Arctic. Marburg and Bigham (2016) found 89% accuracy when classifying benthic mobile megafauna off the Oregon coast. When considering other faunal groups, CV can achieve even higher performances, for example, Siddiqui et al. (2018) classified fish species with up to 96.7% average accuracy.

Even at their best performances, our classifiers would misclassify more than one out of 5 observations if they were used to make novel predictions. This is not good enough to be considered a suitable replacement for manual annotation. To be the tool benthic ecologists need, average performances need to be increased by at least 10 or 15%.

Impact of the number of images in training on performances

In our study, average performance measured as both sensitivity and precision increased with the number of images used in training. Performances obtained with 1000 training images are significantly better than that obtained with fewer images but only marginally so than those obtained with five times less (200) images.

Extrapolation of the data suggests that performances may never greatly exceed those obtained with 1000 training images regardless of how many images are used.

It has been generally demonstrated that more data is preferable when modelling (Enric et al. 2013) and training classifiers (Lu & Weng 2007, Maxwell et al. 2018). Unsurprisingly then, our results suggest that the number of training images has a clear positive effect on performance, particularly on sensitivity. Sun et al. (2017) tested their generalist object classifiers with 10, 30 and 100 million images and observed a clear increase in performance. Siddiqui et al. (2018) also found that increasing the size of a dataset by 25% (20000 to 25000 images) resulted in a 6.6% increase in the accuracy of the same CNN.

More data, however, is not a simple solution to low performances as the relationship between the amount of training data and performance is not linear. Sun et al. (2017) report a logarithmic relationship between the size of the training set and performance. These authors gained less than 20% increase in performance by adding 90 million images to their training set. This logarithmic relationship has also been reported by Favret & Sieracki (2016) in their fly species classifiers. These authors note a diminishing return of adding more training data and observed little gain when doubling their training size from 50 to 100 images. Cho et al. (2015), who classified computed tomography images of six human body parts, found the same logarithmic relationship and, although it was 95.7% with 200 training images, their desired 99.5% accuracy target was only reached with 4092 images. Thus, there is an optimal size to every dataset and a point beyond which more training data results in very little gain. This point can be determined by the goal of the study and what is considered acceptable performance. With our methodology, this point occurs at 200 images, and represents a reasonable amount of manual work for ecologists aiming to build the dataset to train a CNN.

Impact of the number of OTU classes in training on performances

We observed that classifiers with a small (7) number of classes had better performances than those trained with 27 or 52. The difference in performance between the latter two was marginal, although significant.

The number of classes in machine learning studies is usually driven by the dataset and the research question rather than maximizing performance by limiting the number of classes. Thus, few studies have assessed the effect of that number on their performance. Accuracies in the 24 CV-based animal identification studies cited by Favret and Sieracki (2016) and Weinstein (2018) were not significantly correlated to the number of classes used in each classifier. In their large dataset experiment, Sun et al. (2017) also found no difference when training with 1000 or 18000 classes. But in contrast, Favret and Sieracki (2016) observed a counterintuitive increase in performance as more insect species were included into their training set. They hypothesised that, although a higher number of possible outcomes could increase confusion, the higher number of comparison points helped determine the important features of each category. Further tests are needed to disentangle the effect of the number of classes in training or the relative difference in morphology of these classes on performance. In general, practical applications of CV in ecology would benefit from more information on this effect.

Potential application of CV to a real ecological dataset

To deploy classifiers such as these in a “real-life” ecological study, reasonable performances must be achieved while retaining time and cost effectiveness of building the training set.

In our study, no classifier achieved average performance above 0.78, which would mean one misidentification out of 5 predictions, at best. We also observed high interclass variability as some OTU were consistently well identified while others were, on the contrary, always misclassified. Even if the measured average performances were considered acceptable, it would introduce completely false appreciation of the distribution of some OTUs and local diversity.

This variability in both expert and machine classification performance between classes or taxa has been observed by other authors (Beijbom et al. 2015, Cho et al. 2015). Experts in Durden et al. (2016) had various annotation successes for different taxa and Schoening et al. (2012) found that human observers and their semi-automated classifier had variable success at detecting and identifying different taxa but agreed on which one had the best performance. It is therefore sensible to consider the predictions of each OTU class separately and only rely on those for which the classifier achieves good performances, both in precision and sensitivity.

Good performance obtained by our classifier with some specific OTU classes is encouraging and automated annotations could be an appropriate method to study them. The top 10 best and worst OTUs ranked by sensitivity shows that the classifiers are better at identifying large sized organisms exhibiting a low intra-class morphological variability. The majority of the top 10 OTUs were rare (e.g. less than 100 training images). If CV were applied to these rare taxa, there would be a proportionally higher impact of any misidentification or false positive results. Given their relatively low number of occurrences (tens to a few hundreds), a manual verification step (or semi-automated identification), as performed by Schoening et al. (2012) and suggested by Marburg and Bigham (2016), would be easy to perform for a reasonable time investment and ensure the reliability of the predictions. On the other hand, OTU261 and OTU339, both among the top 10 OTU classes, were very abundant in the study area (above 1200 individuals). As manual validation of identification of these OTUs would be impractical, their identification should be fully automated if the classifier is to be deployed on a larger dataset.

With OTU261 and OTU339, high sensitivity (up to 0.95 and 0.92 respectively) and high precision (up to 0.95 and 0.82, respectively) were achieved by the classifiers, meaning they were usually correctly identified and false positives were relatively rare. These performances are equivalent to those of human experts working on a very similar ecosystem (Durden et al. 2016) without the inconsistency over time by individual observers reported by these authors. Therefore, these classifiers could be applied to remaining un-annotated images in our dataset and provide useful presence records of these specific OTUs. This would be a valuable contribution to this study of deep-sea ecosystems.

Classifier A1000 had the best performance of all classifiers and would detect almost all individuals of OTU261 and OTU339, but it needs a large training set, while the A200 classifier has very similar performances but needs five time less training material and is therefore more cost-effective. These group A classifiers however, risk producing a high amount of false positives if they encounter too many individuals of an OTUs they have not been trained on. Thus, it is only applicable if diversity at the study site is low or it is predominantly represented by a small number of OTUs.

These classifiers would not be suitable to survey very diverse ecosystems, like coral reefs.

In the long term, classifiers able to identify as many OTU as possible, even semi-automatically, are undoubtedly more desirable, even if they perform slightly less well. In our study, the C classifiers had marginally lower performances than A, particularly if training with 200 images, but both sensitivity and precision were above 0.9 for OTU261, which is still comparable to manual annotation. Thus, although this design is still valid for identifying specific OTUs, it has the advantage, as it is trained on 52 classes, to be able to automatically identify more OTUs. Even if some of these identifications need to be manually validated, it is more representative of real field studies where many OTUs could be encountered.

Based on our observations on classifier performances, we recommend the following approach to the use of CV in small-scale benthic ecological studies: 1) Build a general classifier to identify OTUs that achieve good performance and quantify the error rate associated with each. This can be an unbalanced design with many OTUs, like group C in the current study. A large number of classes potentially allows more OTUs to be tested. The number of training images should preferably be above 200. 2) Only use the presence prediction of those OTUs that have good performances and regard any other predictions as unknown or absence of those. 3) Consider all remaining OTUs as “unidentified” and leave for manual identification or for later, more efficient, automated classifiers.

Even if the presence records of some OTUs are not sufficient to understand the composition and dynamics of an ecosystem, it will still contribute to it and more importantly, it will take-on some of the annotation time, leaving experts free to perform other tasks while providing provide useful insights in ecology. In the specific case of this study, the automated identification of OTU261 and OTU339 would be useful for deep-sea ecologists, especially if it only requires 200 training images. Indeed, very little is known about the fine scale distribution of these OTUs. *Syringamina fragillissima* (OTU261) is considered habitat forming (Levin et al. 1986, Levin & Thomas 1988) and a Vulnerable Marine Ecosystem under United Nations General Assembly Resolution 61/105 (Assembly 2003). The squat-lobsters *Munida sarsi* or *M. tenuimana* may play an important role in the benthic community as predators or scavengers (Hudson & Wigham 2003) and are suited to examining ecological patterns (Rowden et al. 2010). Extracting the location of these two species from a vast dataset would be a valuable way to study or map their extent and distribution at the basin scale as other studies have done with other faunal groups at fine (Milligan et al. 2016) and broad scale (Rex & Etter 2010, Wei et al. 2010). Besides, this would complement the studies carried out by trawling, which can underestimate diversity of benthic crustaceans (Cartes & Sarda 1992, Ayma et al. 2016) and destroy xenophyophores (Roberts et al. 2000).

This study only deals with the identification of animals and not with their detection on the seabed, which was performed manually. Detection is an essential step in automated image analysis and many solutions have been explored (Cheng & Han 2016, Hollis et al. 2016, Sorensen et al. 2017). A step for object detection needs to be added to the protocol described here to completely automate the process. This study also did not deal with the behaviour of the classifiers when presented with novel OTUs. This situation is unavoidable in real-life ecological datasets, and although methods exist for novelty detection (Pimentel et al. 2014), this remains to be integrated into our methodology.

Conclusion

Our results demonstrate CV based image annotation cannot entirely replace manual annotation of benthic images at present, but that usable results can be obtained for specific taxa with open-source software, very little tuning and optimisation of the CNN itself and a relatively small training dataset (200 images). These results can inform the distribution of these specific taxa in a more robust way than currently possible.

This does not immediately solve the many challenges of benthic ecology but could initiate momentum and catalyse further development of CV based methodology in this area as these tools are becoming more accessible to non-specialists. Indeed, there is still much room left for improving classifier performance with better image pre-processing prior to the training or better tuning of the CNN, and more research could lead to promising methodological development. In the age of big data and global open research, the participation of many different actors of research contributing data (Hampton et al. 2013, Hussey et al. 2015), computing power, and above all, taxonomic and informatics expertise (Weinstein 2018) could be synthesised in the development of CV tools able to take on some of the workload of human researchers and increase the pace at which the oceans are explored and sampled and, ultimately, how they are preserved.

Acknowledgments

The authors would like to acknowledge the officers and crew of RV James Cook as well as the Autosub support team who assisted in the collection of the data. The

DeepLinks project is funded by NERC NE/K011855/1. Piechaud's PhD, Hunt's time and the processing of the data in Google cloud were funded by University of Plymouth. We also would like to thank Kirsty Morris, Erik Simon Lledo and Marcus Shirley for their advice on processing the images prior to analysis.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467

Althaus F, Hill N, Ferrari R, Edwards L, Przeslawski R, Schönberg CH, Stuart-Smith R, Barrett N, Edgar G, Colquhoun J (2015) A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the CATAMI classification scheme. PloS one 10:e0141039

Assembly UNG (2003) Oceans and the Law of the Sea. Report of the Secretary General. A/58/65

Ayma A, Aguzzi J, Canals M, Lastras G, Bahamon N, Mecho A, Company JB (2016) Comparison between ROV video and Agassiz trawl methods for sampling deep water fauna of submarine canyons in the Northwestern Mediterranean Sea with observations on behavioural reactions of target species. Deep Sea Research Part I: Oceanographic Research Papers 114:149-159

Beijbom O, Edmunds PJ, Kline DI, Mitchell BG, Kriegman D Automated annotation of coral reef survey images. Proc Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE

Beijbom O, Edmunds PJ, Roelfsema C, Smith J, Kline DI, Neal BP, Dunlap MJ, Moriarty V, Fan T-Y, Tan C-J (2015) Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. PloS one 10:e0130312

Bicknell AW, Godley BJ, Sheehan EV, Votier SC, Witt MJ (2016) Camera technology for monitoring marine biodiversity and human impact. Frontiers in Ecology and the Environment 14:424-432

Borja A, Elliott M, Snelgrove PVR, Austen MC, Berg T, Cochrane S, Carstensen J, Danovaro R, Greenstreet S, Heiskanen A-S, Lynam CP, Mea M, Newton A, Patrício J, Uusitalo L, Uyarra MC, Wilson C (2016) Bridging the Gap between Policy and Science in Assessing the Health Status of Marine Ecosystems. Frontiers in Marine Science 3

Brandt A, Gutt J, Hildebrandt M, Pawlowski J, Schwendner J, Soltwedel T, Thomsen L (2016) Cutting the Umbilical: New Technological Perspectives in Benthic Deep-Sea Research. Journal of Marine Science and Engineering 4:36

Bullimore RD, Foster NL, Howell KL (2013) Coral-characterized benthic assemblages of the deep Northeast Atlantic: defining "Coral Gardens" to support future habitat mapping efforts. ICES J Mar Sci 70:511-522

Cartes JE, Sarda F (1992) Abundance and diversity of decapod crustaceans in the deep-catalan sea (Western mediterranean). Journal of Natural History 26:1305-1323

Cheng G, Han J (2016) A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 117:11-28

Cho J, Lee K, Shin E, Choy G, Do S (2015) How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv preprint arXiv:151106348

Clark MR, Consalvey M, Rowden AA (2016) Biological Sampling in the Deep Sea. John Wiley & Sons

Costello MJ, Coll M, Danovaro R, Halpin P, Ojaveer H, Miloslavich P (2010) A census of marine biodiversity knowledge, resources, and future challenges. PloS one 5:e12110

Culverhouse PF, Macleod N, Williams R, Benfield MC, Lopes RM, Picheral M (2014) An empirical assessment of the consistency of taxonomic identifications. Marine Biology Research 10:73-84

Culverhouse PF, Williams R, Reguera B, Herry V, González-Gil S (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Marine Ecology Progress Series 247:17-25

Danovaro R, Aguzzi J, Fanelli E, Billett D, Gjerde K, Jamieson A, Ramirez-Llodra E, Smith CR, Snelgrove PVR, Thomsen L, Dover CLV (2017) An ecosystem-based deep-ocean strategy. Science 355:452-454

Durden JM, Bett BJ, Schoening T, Morris KJ, Nattkemper TW, Ruhl HA (2016) Comparison of image annotation data generated by multiple investigators for benthic ecology. Marine Ecology Progress Series

Edgington DR, Cline DE, Davis D, Kerkez I, Mariette J Detecting, tracking and classifying animals in underwater video. Proc OCEANS 2006. IEEE

Enric JdF, David M, Foster P (2013) Predictive Modeling With Big Data: Is Bigger Really Better? Big Data 1:215-226

Favret C, Sieracki JM (2016) Machine vision automated species identification scaled towards production levels. Systematic Entomology 41:133-143

Gaston KJ, O'Neill MA (2004) Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences 359:655-667

Gomes-Pereira JN, Auger V, Beisiegel K, Benjamin R, Bergmann M, Bowden D, Buhl-Mortensen P, De Leo FC, Dionísio G, Durden JM (2016) Current and future trends in marine image annotation software. Progress in Oceanography

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH (2013) Big data and the future of ecology. Frontiers in Ecology and the Environment 11:156-162

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773-785

Hollis DJ, Edgington D, Cline D (2016) Automated Detection of Deep-Sea Animals. Digital Commons, http://digitalcommons calpoly edu/star/370

Howell KL, Davies JS (2016) Deep-sea species image catalogue, On-line version 2. https://deepseacruorg/2016/12/16/deep-sea-species-image-catalogue/

Hudson IR, Wigham BD (2003) In situ observations of predatory feeding behaviour of the galatheid squat lobster Munida sarsi using a remotely operated vehicle. J Mar Biol Assoc Uk 83:463-464

Hussey NE, Kessel ST, Aarestrup K, Cooke SJ, Cowley PD, Fisk AT, Harcourt RG, Holland KN, Iverson SJ, Kocik JF, Mills Flemming JE, Whoriskey FG (2015) Aquatic animal telemetry: A panoramic window into the underwater world. Science 348

Jeffries H, Berman M, Poularikas A, Katsinis C, Melas I, Sherman K, Bivins L (1984) Automated sizing, counting and identification of zooplankton by pattern recognition. Marine biology 78:329-334

Jongman RHG (2013) Biodiversity observation from local to global. Ecological Indicators 33:1-4

Krizhevsky A, Sutskever I, Hinton GE Imagenet classification with deep convolutional neural networks. Proc Advances in neural information processing systems

Langenkämper D, Zurowietz M, Schoening T, Nattkemper TW (2017) BIIGLE 2.0 - Browsing and Annotating Large Marine Image Collections. Frontiers in Marine Science 4

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436

Levin LA, DeMaster DJ, McCann LD, Thomas CL (1986) Effect of giant protozoans (class:Xenophyophorea) on deep-seamount benthos. Marine Ecology Progress Series 29:99-104

Levin LA, Thomas CL (1988) THE ECOLOGY OF XENOPHYOPHORES (PROTISTA) ON EASTERN PACIFIC SEAMOUNTS. Deep-Sea Research Part a-Oceanographic Research Papers 35:2003-2027

Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing 28:823-870

Lucieer VL, Forrest AL (2016) Emerging Mapping Techniques for Autonomous Underwater Vehicles (AUVs). In: Finkl CW, Makowski C (eds) Seafloor Mapping along Continental Shelves: Research and Techniques for Visualizing Benthic Environments. Springer International Publishing, Cham

MacLeod N, Benfield M, Culverhouse P (2010) Time to automate identification. Nature 467:154-155

Mallet D, Pelletier D (2014) Underwater video techniques for observing coastal marine biodiversity: A review of sixty years of publications (1952–2012). Fisheries Research 154:44-62

Manderson T, Li J, Dudek N, Meger D, Dudek G (2017) Robotic Coral Reef Health Assessment Using Automated Image Analysis. Journal of Field Robotics 34:170-187

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of applied Ecology 38:921-931

Marburg A, Bigham K Deep learning for benthic fauna identification. Proc OCEANS 2016 MTS/IEEE Monterey. IEEE

Maxwell AE, Warner TA, Fang F (2018) Implementation of machine-learning classification in remote sensing: an applied review. International Journal of Remote Sensing 39:2784-2817

McClain CR, Rex MA (2015) Toward a Conceptual Understanding of β-Diversity in the Deep-Sea Benthos. Annual Review of Ecology, Evolution, and Systematics 46:623-642

Milligan RJ, Morris KJ, Bett BJ, Durden JM, Jones DOB, Robert K, Ruhl HA, Bailey DM (2016) High resolution study of the spatial distributions of abyssal fishes by autonomous underwater vehicle. Scientific Reports 6:26095

Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, Clune J (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences:201719367

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22:1345-1359

Pimentel MAF, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty detection. Signal Processing 99:215-249

Rampasek L, Goldenberg A (2016) TensorFlow: Biology's Gateway to Deep Learning? Cell Syst 2:12-14

Rex MA, Etter RJ (2010) Deep-sea biodiversity: pattern and scale. Harvard University Press

Ripley B, Venables W, Ripley MB (2016) Package ‘nnet’. R package version:7-3

Roberts JM, Harvey SM, Lamont PA, Gage JD, Humphery JD (2000) Seabed photography, environmental assessment and evidence for deep-water trawling on the continental margin west of the Hebrides. Hydrobiologia 441:173-183

Rohlf FJ, Sokal RR (1967) Taxonomic structure from randomly and systematically scanned biological images. Systematic Zoology 16:246-260

Romero-Ramirez A, Grémare A, Bernard G, Pascal L, Maire O, Duchêne JC (2016) Development and validation of a video analysis software for marine benthic applications. Journal of Marine Systems 162:4-17

Rowden AA, Schnabel KE, Schlacher TA, Macpherson E, Ahyong ST, de Forges BR (2010) Squat lobster assemblages on seamounts differ from some, but not all, deep-sea habitats of comparable depth. Marine Ecology-an Evolutionary Perspective 31:63-83

Schneider S, Taylor GW, Kremer SC (2018) Deep Learning Object Detection Methods for Ecological Camera Trap Data. arXiv preprint arXiv:180310842

Schoening T, Bergmann M, Ontrup J, Taylor J, Dannheim J, Gutt J, Purser A, Nattkemper TW (2012) Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. PloS one 7:e38179

Schoening T, Durden J, Preuss I, Albu AB, Purser A, De Smet B, Dominguez-Carrió C, Yesson C, de Jonge D, Lindsay D (2017) Report on the Marine Imaging Workshop 2017. Research Ideas and Outcomes 3:e13820

Siddiqui SA, Salman A, Malik MI, Shafait F, Mian A, Shortis MR, Harvey ES, Handling editor: Howard B (2018) Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. ICES J Mar Sci 75:374-389

Solan M, Germano JD, Rhoads DC, Smith C, Michaud E, Parry D, Wenzhöfer F, Kennedy B, Henriques C, Battle E, Carey D, Iocco L, Valente R, Watson J, Rosenberg R (2003) Towards a greater understanding of pattern, scale and process in marine benthic systems: a picture is worth a thousand worms. Journal of Experimental Marine Biology and Ecology 285-286:313-338

Sorensen S, Treible W, Hsu L, Wang X, Mahoney AR, Zitterbart DP, Kambhamettu C Deep Learning for Polar Bear Detection. Proc Scandinavian Conference on Image Analysis. Springer

Sun C, Shrivastava A, Singh S, Gupta A Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. Proc 2017 IEEE International Conference on Computer Vision (ICCV)

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z Rethinking the inception architecture for computer vision. Proc Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Team RC (2014) R: A Language and Environment for Statistical Computing. Vienna; 2014.

Van Dover C, Aronson J, Pendleton L, Smith S, Arnaud-Haond S, Moreno-Mateos D, Barbier E, Billett D, Bowers K, Danovaro R (2014) Ecological restoration in the deep sea: Desiderata. Marine Policy 44:98-106

Wei C-L, Rowe GT, Escobar-Briones E, Boetius A, Soltwedel T, Caley MJ, Soliman Y, Huettmann F, Qu F, Yu Z (2010) Global patterns and predictions of seafloor biomass using random forests. PLoS One 5:e15323

Weinstein BG (2018) A computer vision for animal ecology. Journal of Animal Ecology 87:533-545

Wheeler B, Torchiano M (2010) ImPerm: Permutation tests for linear models. R package version 1

Wickham H (2017) Tidyverse: Easily install and load'tidyverse'packages. R package version 1

Williams SB, Pizarro O, Steinberg DM, Friedman A, Bryson M (2016) Reflections on a decade of autonomous underwater vehicles operations for marine survey at the Australian Centre for Field Robotics. Annual Reviews in Control 42:158-165

Wynn R, Bett B, Evans A, Griffiths G, Huvenne V, Jones A, Palmer M, Dove D, Howe J, Boyd T (2012) Investigating the feasibility of utilizing AUV and Glider technology for mapping and monitoring of the UK MPA network. Southampton: National Oceanography Centre

Tables

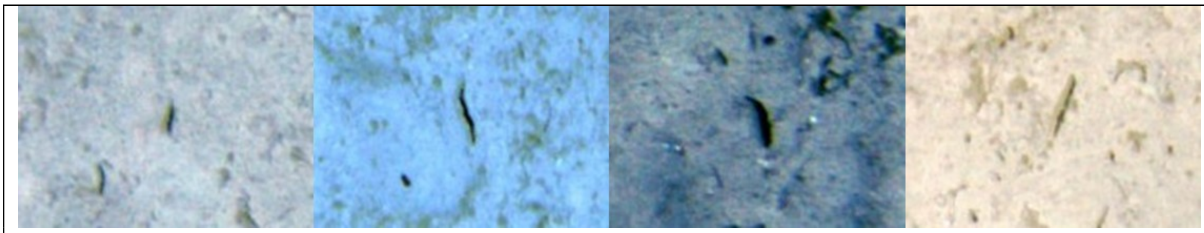
*Table 1: Nomenclature of classifiers names and characteristics. The different classifiers names are a combination of group name and image numbers per Operational Taxonomical Units (OTU) in training. Groups are defined by the number of different OTUs (or classes) in the training set. In the different groups, the OTUs used are those for which the minimum number of images indicated are available. Within each group, treatments refer to the number of images of each class in training. The same treatments (20, 50, 100, 200, 500 and 1000 images per OTU in training) were applied to each group but only the classifiers names **in bold** are balanced (equal number of images for every class). In unbalanced designs, the maximum number of available images is used and is therefore different for each OTU.*

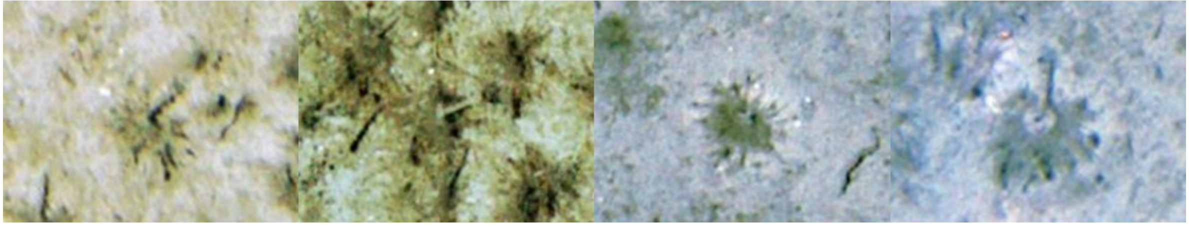
	Groups		
	A	B	C
Number of classes	7	27	52
Minimum number of images available for the OTU to be in the group	1000	100	20
Classifiers names in group **(balanced classifiers in bold)**	**A20, A50, A100, A200, A500, A1000**	**B20, B50, B100**, B200, B500, B1000	**C20**, C50, C100, C200, C500, C1000

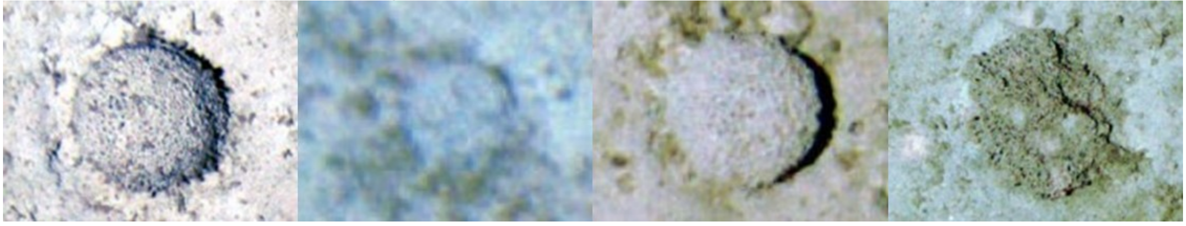
Table 2: Possible outcomes of the classifiers. It indicates how the classifiers predictions compare to the manual annotation (the labels) and if it identifies the Operational Taxonomical Unit (OTU) present on an image correctly.

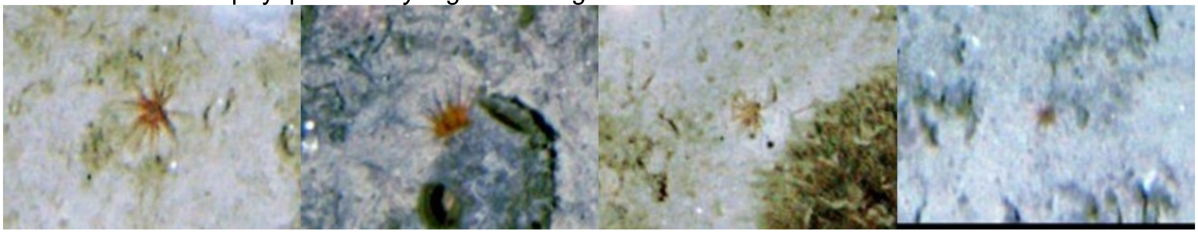
Outcome	**Description**
True Positives	Label is OTU and class predicted is OTU ► Classifier correctly identified the OTU
True Negatives	Label is not OTU and class predicted is not OTU ► Classifier correctly recognized the OTU is not in the image
False Negatives	Label is OTU but class predicted is not OTU ► Classifier misidentified the OTU
False Positives	Label is not OTU but class predicted is OTU ► Classifier misidentified another OTU

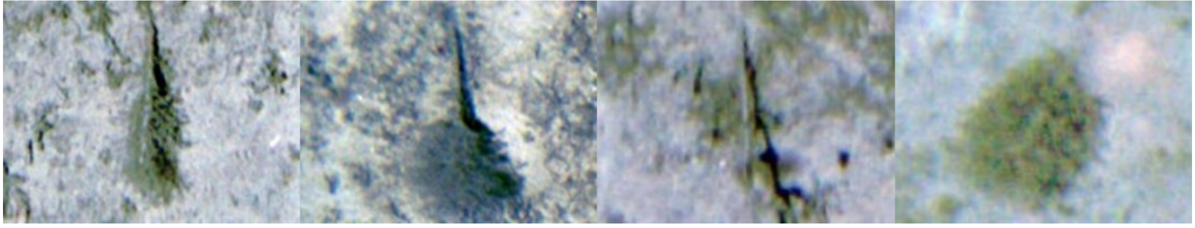
Figures

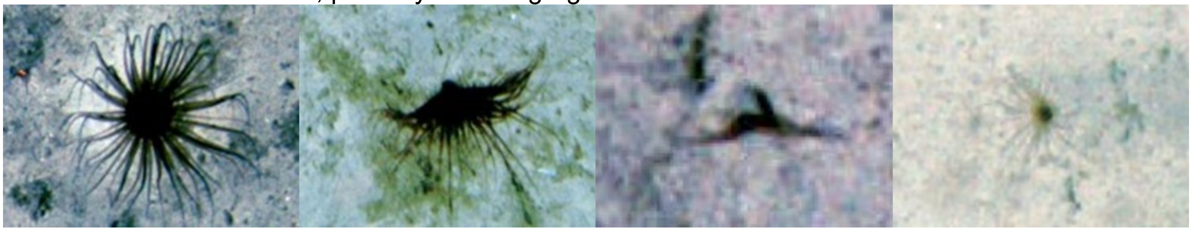

OTU603: Very small elongated sponge. Shape is constant.

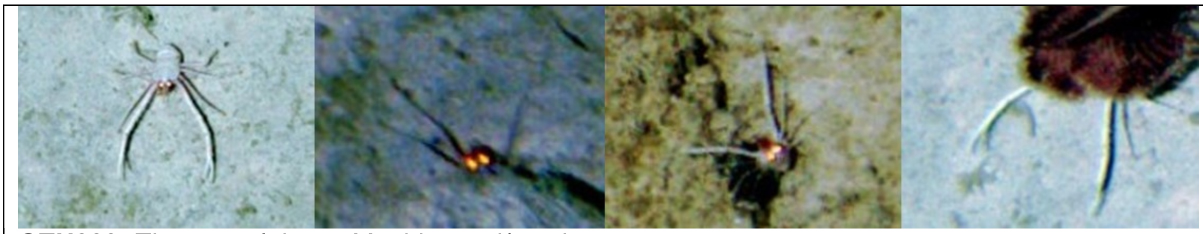

OTU375: Small tube worm. The gills can hide the tube


OTU261: The xenophyophores *Syringamina fragillissima*


OTU23: Small halcampid/edwardsiid anemone


OTU995: Unknown animal, possibly a chrisogorgid


OTU2: Cerianthid anemone of various size



OTU339: The squat lobster *Munida sarsi/tenuimana*

Figure 1: Example images and description of OTUs abundant enough to be in group A. Scale varies. OTUs are ordered by abundance in the original dataset.

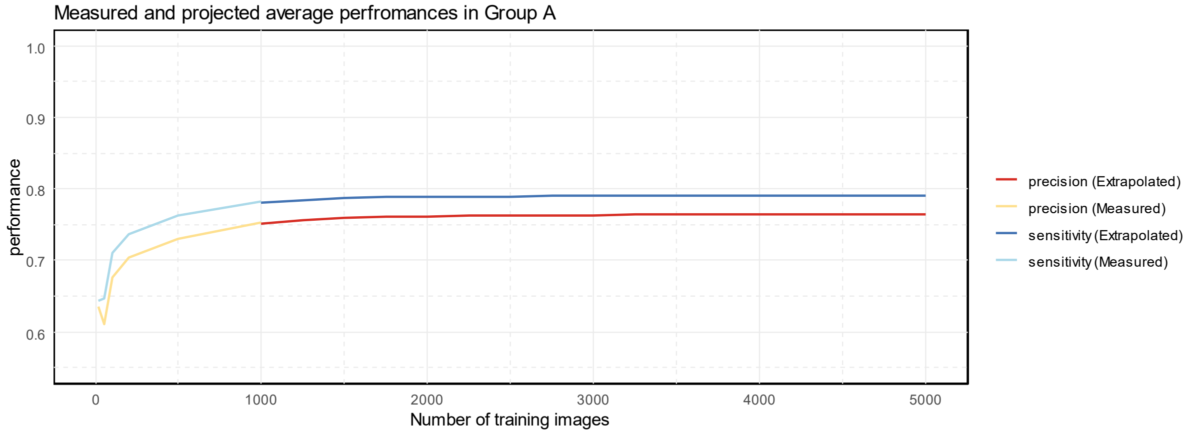
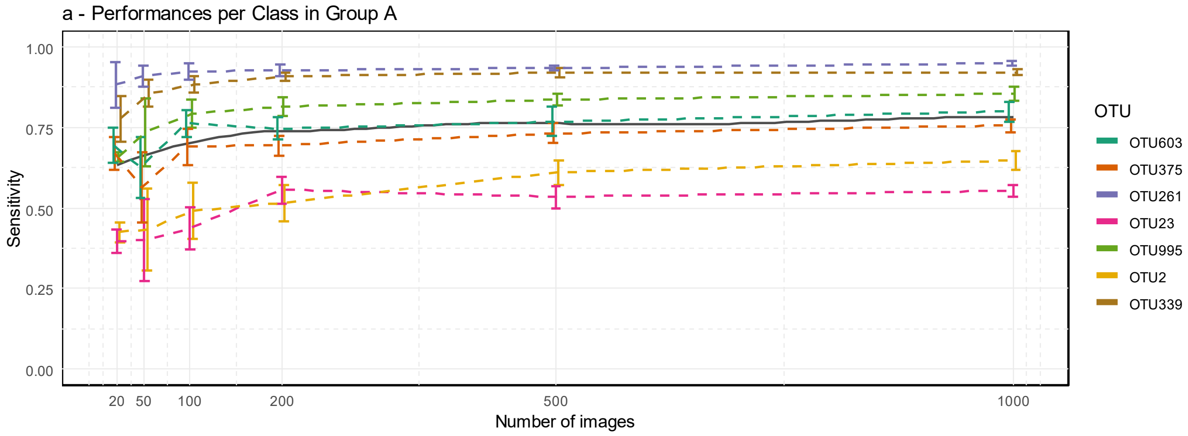


Figure 2: Classifier performances (sensitivity and precision) per number of training images measured (20 – 1000) and extrapolated (1000 – 10000). Grey dots show averaged values across all OTUs for each classifiers



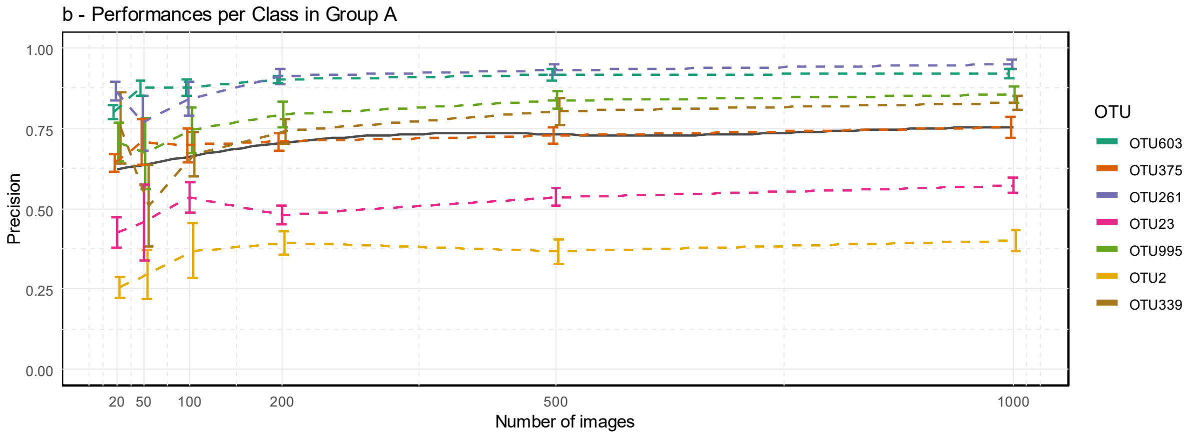


Figure 3: a) Evolution of Sensitivity in Group A classifier trained with an increasing number of images. b) Differences in Precision in Group A classifier trained with an increasing number of images. The black line is 'loess' smoothed curve of the average of all the classes and greyed area is a t-based approximation of the standard error.

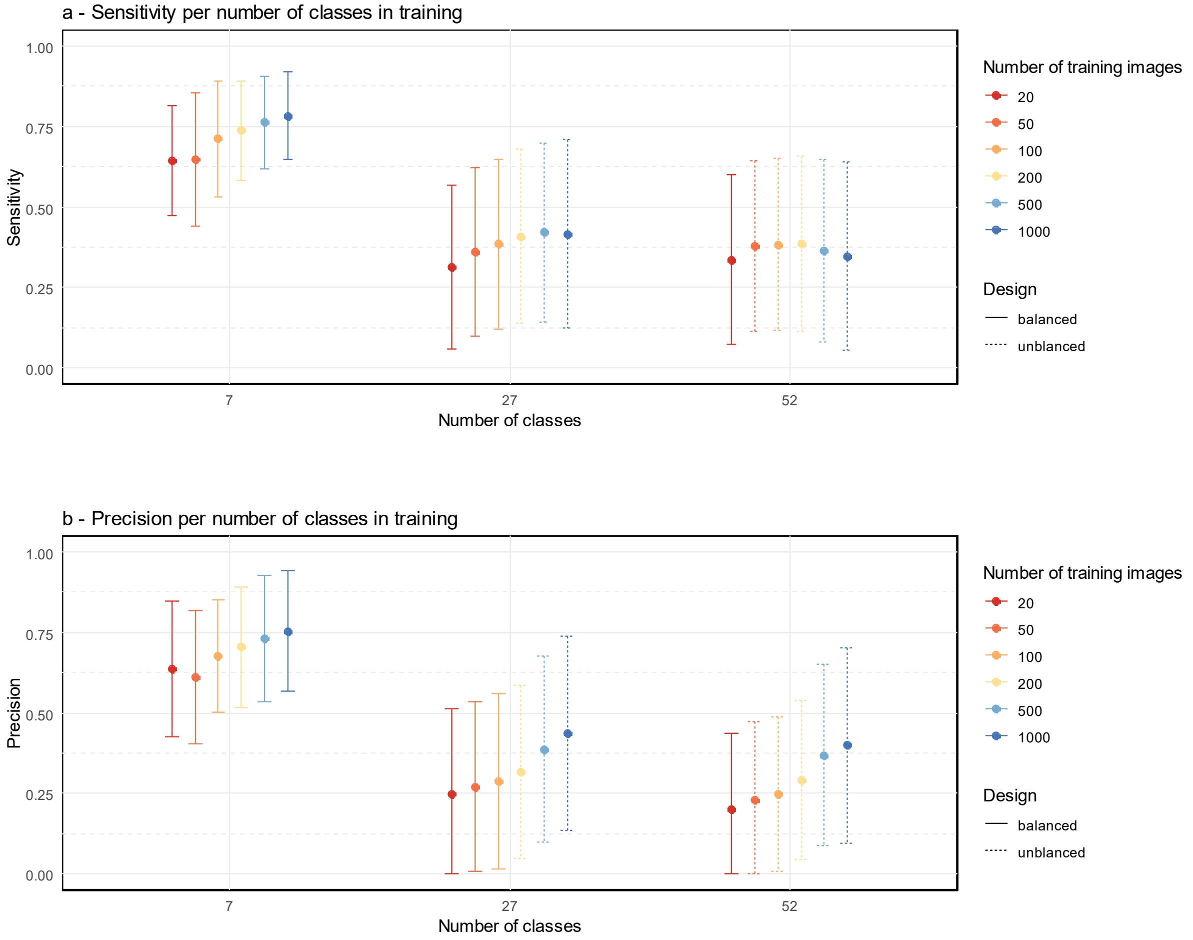


Figure 4: a) Differences in sensitivity in classifiers trained with different number of classes and images. b) Differences in precision in classifier trained with different number of classes and images. Error bars are standard deviation of the 10 random splits.

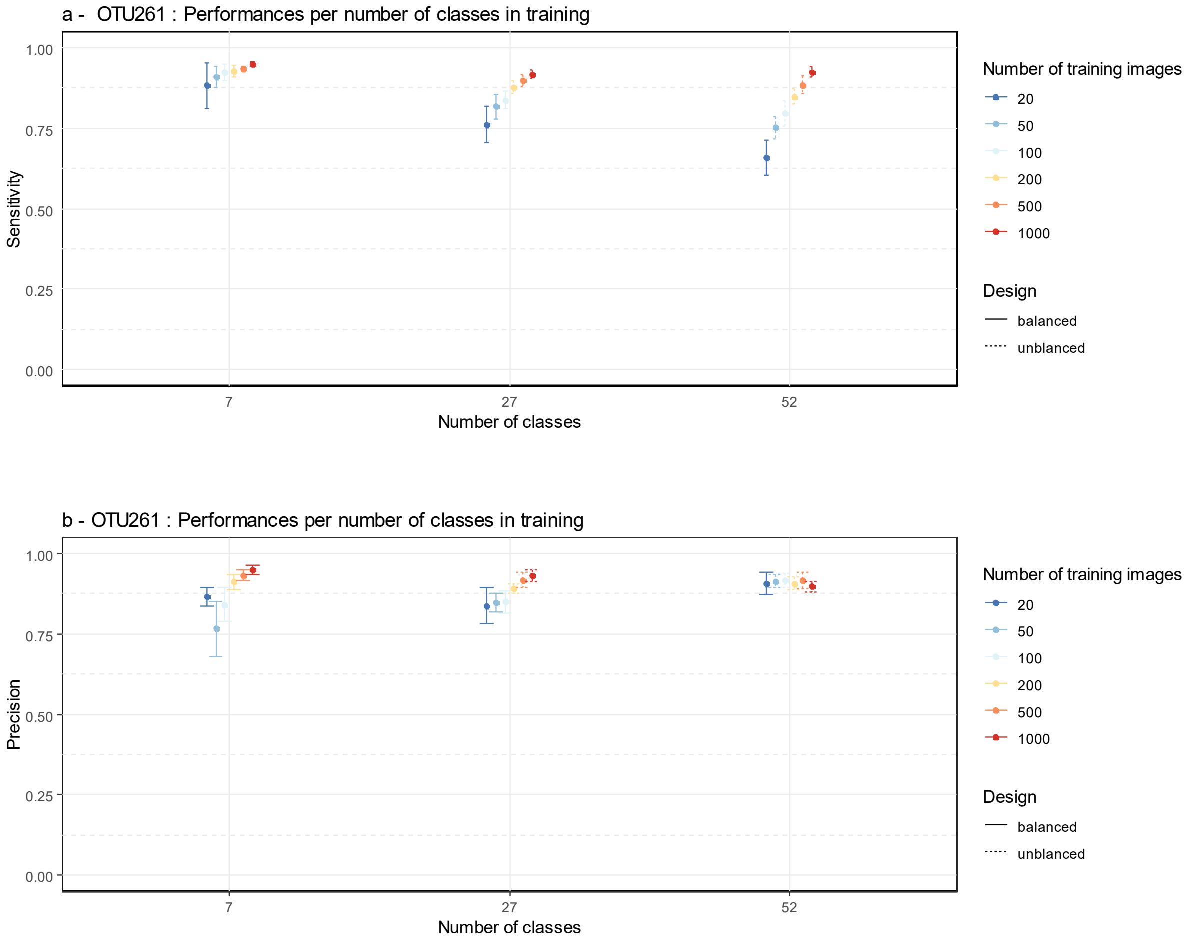


Figure 5: a) Differences in sensitivity for OTU261 in classifier trained with different number of classes and images. Error bars are standard deviation calculated from the 10 random splits. b) Differences in precision for OTU261 in classifier trained with different number of classes and images. Error bars are standard deviation of the 10 random splits

Appendix

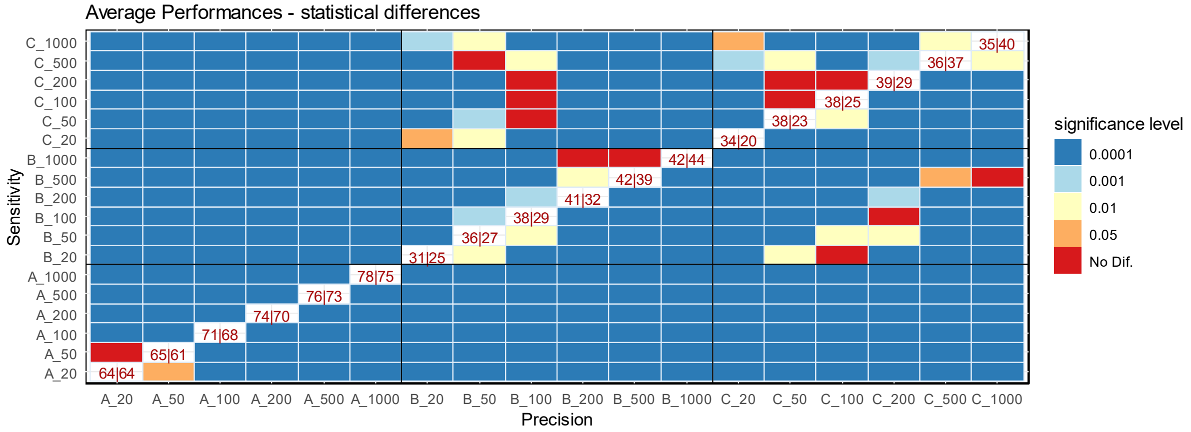


Figure A1: Pair-wise permutation-based analysis of variance of differences in sensitivity (upper left) and precision (lower right) between each treatment. The numbers in central cells indicates sensitivity (left) and precision (right) of corresponding treatments on the axis. Significance level indicate at which alpha threshold the two treatments are significantly different in percentages of maximal value (i.e. 1). No dif. indicates a p-value above 0.05.

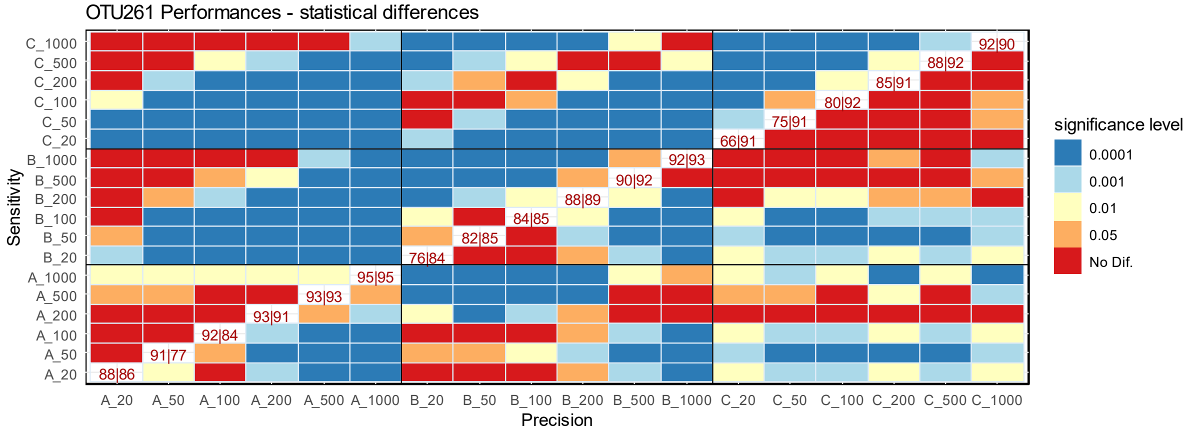


Figure A2: Pair-wise permutation-based analysis of variance of differences in sensitivity (upper left) and precision (lower right) between each treatment. The numbers in central cells indicates sensitivity (left) and precision (right) of corresponding treatments on the axis in percentages of maximal value (i.e. 1). Significance level indicate at which alpha threshold the two treatments are significantly different. No dif. indicates a p-value above 0.05.

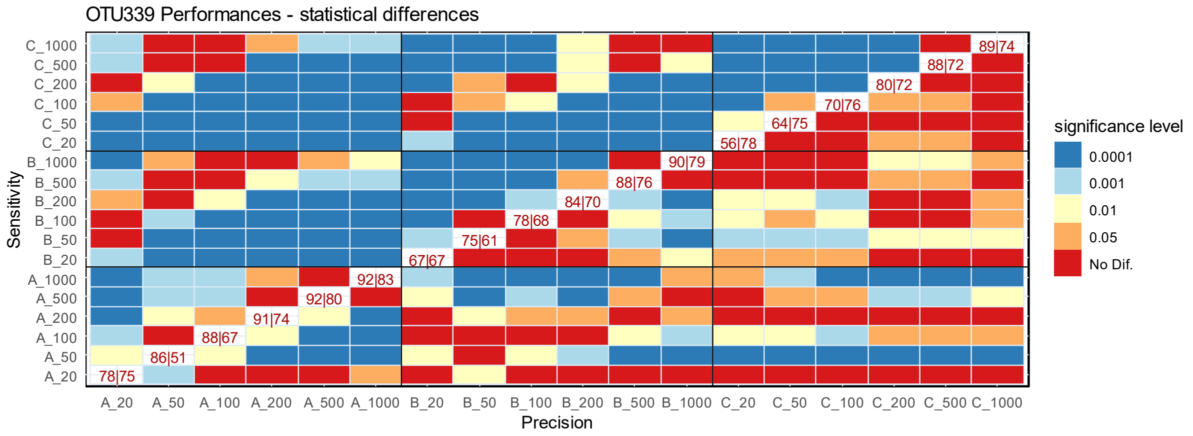
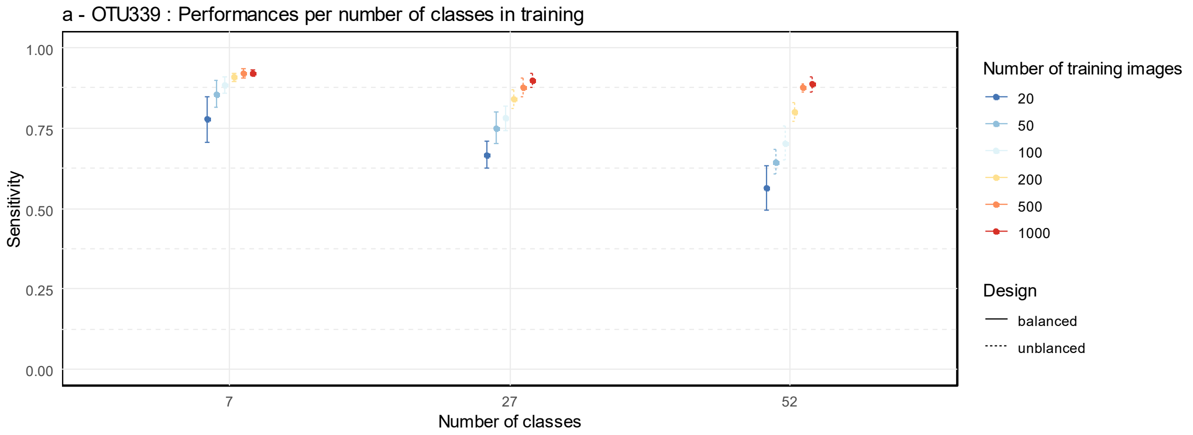


Figure A3: Pair-wise permutation-based analysis of variance of differences in sensitivity (upper left) and precision (lower right) between each treatment. The numbers in central cells indicates sensitivity (left) and precision (right) of corresponding treatments on the axis. Significance level indicate at which alpha threshold the two treatments are significantly different. No dif. indicates a p-value above 0.05.



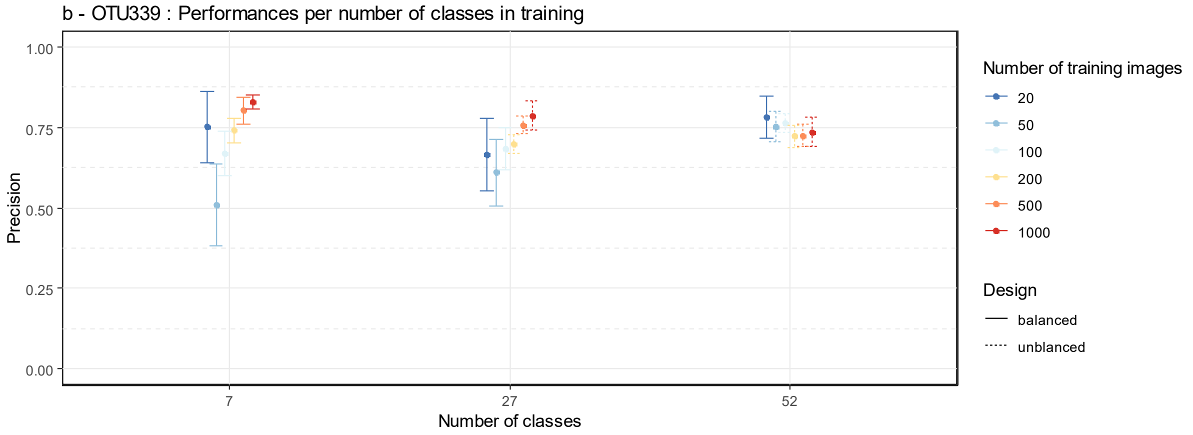


Figure A4 a) Differences in sensitivity for OTU 339 in classifiers trained with different number of classes and images. Error bars are standard deviation calculated from the 10 random splits. b) Differences in precision for OTU 339 in classifier trained with different number of classes and images. Error bars are standard deviation of the 10 random splits.