2019-04-18

Automated identification of benthic epifauna with computer vision

Piechaud, N

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

Automated identification of benthic epifauna with computer vision

3 **Running head:** Applying computer vision to benthic ecology

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

5 Kerry L. Howell[1]

6 1: School of Biological and Marine Sciences, University of Plymouth, Plymouth, PL4

7 8AA, UK

8 2: Controlled frenzy LTD, THINQTANQ, Fairbairn House, Higher Lane, Plymouth,

9 PL1 2AN

10 3: School of Computing, Electronics and Mathematics, University of Plymouth,

11 Plymouth, PL4 8AA, UK

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

13 **Abstract:** Benthic ecosystems are chronically undersampled, particularly in

14 environments >50m. Yet, a rising level of anthropogenic threats makes data

15 collection ever more urgent. Currently, modern underwater sampling tools,

16 particularly Autonomous Underwater Vehicles (AUV), are able to collect vast image

17 datasets, but cannot bypass the bottleneck formed by manual image annotation.

18 Computer Vision (CV) offers a faster, more consistent, cost effective and a sharable

19 alternative to manual annotation. We used Tensorflow to evaluate the performance

20 of the Inception V3 model with different numbers of training images, as well as

21 assessing how many different classes (taxa) it could distinguish. Classifiers (models)

22 were trained with increasing amounts of data (20 to 1000 images of each taxa) and

1

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

2

48 collected, stored, and processed at a time. Since the 19th century, various

49 technological innovations have attempted to bypass this bottleneck.

50 Benthic ecosystems are traditionally sampled by trawls, cores and other physical

51 means. These physical samples are costly to collect and process, and logistically

52 challenging to store (Clark et al. 2016). While physical samples remain the mainstay

53 of benthic surveys, use of underwater imaging technologies is increasingly popular

54 among marine ecologists (Solan et al. 2003, Bicknell et al. 2016, Brandt et al. 2016,

55 Romero-Ramirez et al. 2016). These technologies offer a less invasive, more cost

56 effective method of survey, and storage space for image data is virtually unlimited

57 (Mallet & Pelletier 2014). Underwater imaging is now regularly utilised alongside

58 other sampling tools to provide a comprehensive view of the marine environment.

59 Modern underwater sampling vehicles, and particularly Autonomous Underwater

60 Vehicles (AUV), have great potential in providing the step-change in the rate of data

61 gathering that is needed to support sustainable marine environmental management.

62 They are capable of collecting large numbers of images of the sea bed in a single

63 deployment (Lucieer & Forrest 2016, Williams et al. 2016). For example, a 22 hour

64 AUV dive can deliver more than 150,000 images of the seafloor along with other

65 types of environmental data (Wynn et al. 2012). Comparatively, trawls and Remotely

66 Operated Vehicles (ROV) cover less ground per dive and the ship and its crew are

67 unable to operate any other benthic equipment while they are deployed (Brandt et al.

68 2016, Clark et al. 2016).

69 To translate the information contained in images into semantic data that can then be

70 used in statistical analysis, a step of manual analysis (or annotation) is conducted by

71 trained scientists. Human observers, even highly-trained, do not achieve 100%

3

72   correct classification rates and are highly inconsistent across time and across

73   annotators (Culverhouse et al. 2003, Culverhouse et al. 2014, Beijbom et al. 2015,

74   Durden et al. 2016). Besides, manual image annotation results are subject to

75   observer bias, meaning interpretations vary depending on the annotators experience

76   and their mood changes across the analysis process (tiredness, boredom or stress,

77   etc…) (Culverhouse et al. 2003, Durden et al. 2016). The results (format, taxonomic

78   resolution and nomenclature) of these analyses also tend to differ from one

79   institution, project or individual annotator to another. This lack of standardisation

80   makes merging and comparing datasets difficult (Bullimore et al. 2013, Althaus et al.

81   2015, McClain & Rex 2015), and the data quality is not always consistent. More

82   importantly, manual analysis is a time consuming process, which forms the current

83   bottleneck in image based marine ecological sampling (Edgington et al. 2006,

84   Beijbom et al. 2015, Schoening et al. 2017). The growing trend towards use of AUVs

85   for seafloor biological survey will only worsen this situation.

86   Artificial intelligence (AI) and computer vision (CV) provide potential means by which

87   to both accelerate and standardise the interpretation of image data (Culverhouse et

88   al. 2003, MacLeod et al. 2010, Beijbom et al. 2012, Favret & Sieracki 2016).

89   Although using AI for biological research has a long history (Rohlf & Sokal 1967,

90   Jeffries et al. 1984, Gaston & O'Neill 2004), it has always been challenging to

91   implement for non-specialists and requires skills and materials that most biologists

92   do not have access to (Gaston & O'Neill 2004, Rampasek & Goldenberg 2016).

93   CV has been successfully applied to benthic species identification by a growing

94   number of studies (Edgington et al. 2006, Beijbom et al. 2015, Marburg & Bigham

95   2016, Manderson et al. 2017, Norouzzadeh et al. 2018, Schneider et al. 2018) but

96   has yet to be made into an easy to use tool that any biologist in the field can

4

97   implement as an alternative to manual image annotation and integrate with previous

98   analysis. Multiple potential commercial applications, the availability of new tools as

99   open software, as well as the improvement of hardware capacity are driving new

100 developments in AI (e.g. neural networks and deep learning). This is likely to change

101 how AI can be employed in the field of scientific research (Rampasek & Goldenberg

102 2016, Weinstein 2018). In parallel, new image analysis and data science software

103 allow an easier and more efficient integration of various tools into the research

104 process, from data collection to final scientific or public outreach material (Gomes-

105 Pereira et al. 2016). These new technologies are potentially enabling full automation

106 of the annotation process and could revolutionise ecological research (Weinstein

107 2018).

108 While the principle of automated classification (automated assignation of pre-

109 established classes to objects on images) has been validated, few practical

110 examples exist of AI-based methods used to identify benthic animals from images

111 acquired by AUV. Consequently, implementing an automated species classifier is a

112 potentially time consuming investment for an uncertain return. Relying on proven

113 manual methods remains the safe option for researchers. Practical guidance is

114 needed to help ecologists decide whether adopting AI and CV is feasible and would

115 fit their dataset and scientific objectives.

116 To make that decision, benthic ecologists need to know:

117 • What level of accuracy and uncertainty can be expected from CV annotation
118 and does it match or approximate the accuracy of human annotators.

119 • How much material is needed to train a classifier and is a limited amount
120 obtained from a single study sufficient.

5

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

122 In this study, we investigate these issues by using an open access algorithm to build

123 a Convolutional Neural Network (CNN) to identify benthic animals in seafloor

124 images, obtained from a single deployment of the UK's Autosub6000 AUV.

125 Technically speaking, we seek to train an automated classifier that is able to

126 determine which taxa an animal on an image most likely belongs to, using a list of

127 pre-defined taxa (or classes). Specifically we ask, 1) what impact does the number of

128 images, on which the classifier is trained, have on its performance? and 2) What

129 impact does the number of classes, on which the classifier is trained, have on its

130 performance? In addition, we provide a case study in the application of CV to an

131 unbalanced ecological dataset.

Method

132

Study area and data collection:

133

134 All the images used in this study were collected by the UK's national AUV

135 Autosub6000 in May 2016 as part of the NERC funded DeepLinks (JC136) research

136 cruise. The images were taken as part of a 1880 m long transect at station 26 of that

137 cruise at 1200 meters depth on the north-east side of Rockall Bank, N.E. Atlantic.

138 This region was selected for the study due to the flat topography and low likelihood

139 of disturbance, making it ideal for AUV deployment. The AUV was equipped with a

140 downward facing Grasshopper2 GS2-GE-50S5C camera from Point Grey Research.

141 The AUV was flown at $1.1ms^{-1}$ speed, at 3m ±0.1 m off bottom and took images

142 every second, resulting in near overlapping image coverage. The surface area of

143 each image is between 1 and 2.5 m^2, and the resolution is 2448 x 2048 at 5 mega

144 pixels.

145 In total, 1165 raw photos of the seabed were manually annotated by a single

146 observer with the Biigle 2.0 software (Langenkämper et al. 2017) using a regional

147 catalogue of Operational Taxonomical Units (OTU) developed (Howell & Davies

148 2016). Within the Biigle 2.0 software, location (X and Y coordinates in pixels within

149 the photo for point annotations, or X, Y and radius for individuals marked using a

150 circle) and identity of individual OTUs annotated within each image was recorded

151 and stored. For each OTU, all individual annotations were visually inspected using

152 the "Largo" evaluation tool in Biigle 2.0, to ensure consistency in identification and

153 reduce error.

154 ## Image data

155 Manual image annotation resulted in a dataset consisting of 41208 individuals

156 belonging to 148 OTUs. Each individual was then cropped from the raw image,

157 together with its assigned OTU label, using a custom Python (www.Python.org)

158 script. For each annotation, a square of 40 pixels or more, positioned manually on X

159 and Y coordinates of the centre of the animal, was fitted and cropped out. For

160 animals bigger than 40 pixels, the size of the square was manually set to encompass

161 the whole individual. These cropped image slices and associated OTU labels (to

162 become classes in the model training design) formed the input used in the CNN.

163 ## Tensorflow and transfer learning

164 Rather than train our own neural network, we used transfer learning (Pan & Yang

165 2010) to retrain the Inception V3 model (Szegedy et al. 2016), a CNN built in the

166 freely available library Tensorflow (Abadi et al. 2016).

167 CNNs are a particular architecture of neural networks, more specifically, deep

168 learning, particularly suited to image analysis (Krizhevsky et al. 2012, LeCun et al.

7

169   2015). A CNN has the capacity to detect and match patterns in images thereby

170   "learning" what features are relevant to differentiate objects and, subsequently,

171   classify them accordingly.

172   Tensorflow (TF) is a C++ based library but has a Python Application Programming

173   Interface (API) that makes it easier to train, tune and deploy neural networks.

174   Transfer learning is a method allowing a CNN built on a large dataset to be re-

175   repurposed into a classifier capable of distinguishing between classes it was not

176   initially trained on. The strength of this method is that the dataset on which it is

177   transferred does not need to be as large as it should be to train a CNN from the

178   beginning. Here, we were able to train a classifier with a tens to hundreds of images

179   per class (in our case, OTUs) instead of millions.

180   **Classifier training and testing**

181   A random 75-25% split was applied to every OTU in order to separate images used

182   for training the classifier and those used for testing. The training and test data sets

183   for all OTUs were then combined into single 'training' and 'test' datasets.

184   The OTUs the classifier was trained to identify are referred to as classes and only

185   those OTUs for which there were a sufficient number of image slices (individual

186   observations) available were selected for use in training. The minimum number of

187   images needed for training was set to 20. This means that for an OTU to be included

188   in the study at least 27 image slices were needed, 20 for training and 7 for testing.

189   Out of the 148 OTUs observed, 52 were above that threshold. The remaining 96

190   OTUs represented 3.19% of the total number of individual annotations and were

191   removed from the dataset.

8

192 The classifier was trained on the training dataset and then predictions were made on

193 the test dataset. For each cropped image slice in the test dataset, TF gave a score

194 for each of the possible OTU classes for which it had been trained. The scores range

195 from 0 to 1 (the sum of scores for all classes being 1) and represent the model's

196 confidence that the slice belongs to the corresponding class. The final prediction was

197 the OTU class that received the highest score. The prediction was then compared to

198 the manually assigned OTU class.

199 To measure the effect of the number of training images (or limit) on the accuracy and

200 confidence of the predictions, the training data set was filtered so each OTU class

201 was represented by 20, 50, 100, 200, 500, and 1000 images (Table 1). A classifier

202 was then trained on each of these six pools of images and tested using the test data

203 set. Only seven OTUs were frequent enough to be used with these six limits (Figure

204 1).

205 The combination of groups and limits is referred to as treatments and designation of

206 each treatment follows the nomenclature in table 1 (e.g. A1000 is group A, limit

207 1000). Each treatment was repeated 10 times with different random splits between

208 testing and training data.

209 To measure the effect of the number of OTU classes used to train the CNN on its

210 capacity to correctly classify the test dataset, we used three training datasets each

211 with different numbers of classes (referred to as groups) (Table 1). The number of

212 classes is defined by the number of available images per OTU so classifiers can be

213 trained on a set number of images for every class while retaining enough images for

214 testing. Group A contained 7 classes for which more than 1000 images was

215 available; group B contained 27 classes for which more than a 100 images were

9

216   available; and group C contained 52 classes for which more than 20 images were

217   available.

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

219   Note that when the limit is above the available number of images, the classes with

220   less images were trained with the maximum number available regardless of the limit.

221   This results in class imbalance in the model training for some treatments in group C

222   with more than 20 images and in B with more than 100 images (balanced treatments

223   are listed in Table 1). To assess the effect of the number of OTU classes used to

224   train the CNN on its capacity to correctly classify the test dataset only balanced

225   designs were used.

226   In total, 180 (3x6x10) classifiers were trained and tested. All the CNNs were trained

227   in the Google Cloud ML (https://cloud.google.com/) remote computing facility.

228   To be applied to a "real life" ecological study, the classifiers have to maximize

229   performances while minimizing the initial effort needed to build the training dataset.

230   To assess appropriate use of CV on a 'real life' dataset we considered all possible

231   combinations of numbers of training image and numbers of OTU classes in an

232   unbalanced design. Average performances and individual OTU performances were

233   assessed.

234

235   **Analysis and performances evaluation**

236   Considering each class, the observation can be a presence (the OTU is present on

237   the image) or an absence (the OTU is not on the image and another OTU is). The

238   different possible outcomes or predictions of the classifier are detailed in Table 2.

10

239   The respective number of each outcome type (the confusion matrix) was used to

240   calculate performance metrics.

241

242   The classification accuracy is the percentage of predictions that are correct

243   (prediction matches observation) and is often used to evaluate performances in ML

244   studies. This measure ignores the differences between classes, thus we used two

245   model evaluation metrics which rely on a confusion matrix (Manel et al. 2001)

246   explained in Table 2.

247   -   **Sensitivity,** also referred to as *true positives rate* or *recall*. It varies between 0

248   and 1. It quantifies the proportion of individuals of a given OTU in the testing

249   set that are correctly identified. A value of 1 means that all individuals of a

250   given OTU are identified as such.

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

251

252   -   **Precision,** *or Positive Predictive Value*. It varies between 0 and 1. It

253   quantifies the proportion of true positives among the individual identified as a

254   given OTU. A value of 1 means all the individual identified as a given OTU

255   class are indeed that OTU.

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

256

257   Average and standard deviation for all metrics were calculated for each class within

258   each treatment and then averaged over other grouping factors. This gave an

11

259 estimation of the overall performance of the classifiers. The performances of the

260 classifiers for each individual class were also carefully analysed.

261 Differences in metrics were statistically tested with a permutation-based analysis of

262 variance in the "lmPerm" package in R (Wheeler & Torchiano 2010). We report p-

263 values classified with five levels of significance: more than 0.05 or non-significant,

264 less than 0.05, less than 0.01, less than 0.001 and less than 0.0001. Relationships

265 between number of images and performance were extrapolated with a neural

266 network regression in the "nnet" package in R (Ripley et al. 2016) projected over

267 1000 to 10000 images. All data analyses were carried out in R (Team 2014) using

268 the "tidyverse" package (Wickham 2017).

269

270 ## Results

271 The results are presented in three sections. First, questions related to the impact of

272 the number of training images are addressed, then the effect of the number of

273 classes in the training set is assessed, and finally the results relevant to choosing the

274 best method in our case study are presented.

275 ### Impact of the number of training images on performance

276 Average performance, measured as both sensitivity and precision, increases with an

277 increasing number of images used (Figure 2). For sensitivity, there is an average

278 increase from 0.64 to 0.78 when moving from 20 to 1000 images, respectively. This

279 is mirrored by increases in precision from 0.63 to 0.75 when moving from 20 to 1000

280 images, respectively. Non-linear extrapolations of average sensitivity and precision

281 show that performances reached with 1000 training images may be close to an

282   asymptote and performances obtained with additional training material probably

283   plateau below 0.78 for sensitivity and 0.75 for precision (Figure 2). This suggests

284   that the model is unable to achieve perfect performance regardless of how many

285   additional images are used in training.

286

287   The number of images has a clear positive effect on performances. For almost all

288   pairs of models compared in (Figure A1), performance values are statistically

289   significantly different ($p < 0.05$) and very often, significance is very high (p-value <

290   0.0001). There are a few exceptions like between A20 and A50 classifiers where p-

291   value > 0.05 for sensitivity and between 0.01-0.05 for precision or the B1000

292   classifier, for which there is no significant difference between B500 and B200 in

293   sensitivity. However, measured difference in performance between sequential

294   models becomes vanishingly small at higher numbers of training images, such that

295   the difference between A200 and A1000 classifiers is 0.04 for sensitivity and 0.05 for

296   precision. This suggests little to no improvement is gained in model performance by

297   using more than 200 training images.

298

299   There are strong between-OTU differences in classifier performances (Figure 3). All

300   classifiers have high sensitivity for OTU261 and OTU339, even the A20 classifier

301   (0.88 and 0.77, respectively). For OTU2 and OTU23, classifiers have more variable

302   and lower sensitivity regardless of the number of training images used. OTU261 is

303   very constant in shape and colour and has a distinctive pattern on its outside.

304   OTU339 can be in different pose or orientation within an image but has a number of

305   distinguishing features such as its reflective eyes, and its long, often spread-out

13

306 limbs. OTU2 and OTU23 are both anemones. OTU2 is a cerianthid (a tube

307 anemone) of various size and orientation and OTU23 is a

308 Halcampidae/Edwardsiidea like anemone of very small size.

309 The OTUs for which precision is highest are not necessarily those for which

310 sensitivity is highest. The highest precision observed was for OTU261 but the

311 second highest precision observed was for OTU603, which has a lower sensitivity.

312 For some classes (OTU261 or OTU339), precision is lower with 50 training images

313 compared to 20 training images.

314

315 **Impact of the number of classes on classifier performance**

316 Classifiers trained with 7 classes (group A) had significantly better sensitivity (Figure

317 A1) and precision than equivalent classifiers trained on more classes but the same

318 number of images (Figure 4). Variability in performance was also lower for classifiers

319 trained with fewer classes. Average sensitivity decreased from 0.71 to 0.38, and

320 average precision decreased from 0.69 to 0.32, when moving from 7 to 27 classes.

321 This suggests a negative effect of the number of classes on performance; however,

322 on average, there is only a minor drop in performance (0.018 in sensitivity and 0.035

323 in precision) between classifiers trained on 27 and 52 classes. Interestingly, B100

324 and C100 both have sensitivity of 0.38 (no statistical difference) and C20 has higher

325 (+ 0.02) sensitivity than B20.

326 OTUs that perform well in a group tend to perform well in others. OTU261 and

327 OTU339 are in the top 10 for each group although their performances are lower in

328 group B and C.

14

329

Application of CV to an unbalanced ecological dataset

331   When considering all treatments in an unbalanced design (Figure 4), the average

332   sensitivity per treatment ranges from 0.32 to 0.78. The highest sensitivity was

333   achieved by the A1000 classifier (7 classes, with 1000 training images in each class)

334   while the lowest was achieved by the B20 and C20 classifiers (27 and 52 classes,

335   respectively, and 20 images in each class). A1000 also had the highest precision

336   (0.75), with the lowest precision observed in the C20 classifier (0.20). Sensitivity of

337   the C1000 classifier (where class imbalance is highest) was lower than in the C100

338   and C200 classifiers but precision simply increases with the number of training

339   images, although this could be an artefact driven by the improvement of precision on

340   the most abundant classes.

341 When considering individual OTUs, performance was unacceptably low for most, but

342 not all as some had sensitivity and precision greater than 0.85. Based on average

343 sensitivity across all treatments, the top 10 and the bottom 10 OTU classes were

344 identified. The top 10 classes were large animals with consistent or distinctive

345 shape, colour and patterning. They were not necessarily the most abundant classes

346 as six of them were only present in group C, for which there are less than 100

347 training images, and only two in A, for which there are at least 1000 training images.

348 The two of these OTU present in group A had better average precision than any

349 other OTU class in the top 10. Those OTU classes with the worst performances are

350 generally those for which there are fewer training images (group C). They also tend

351 to be smaller organisms, have colours similar to the background and have very

352 variable shapes and sizes.

353 In this dataset, CV could be applied to OTU261 and OTU339. These OTUs were

354 both very abundant in the study area, justifying automated annotation, and they both

355 had very high performances, making their identification by the classifier reliable

356 (Figure 5).

357

358 The performance of CV for OTU261 and OTU339 was maximised in the A1000

359 classifier with only 7 classes and 1000 training images. The A200 classifier also

360 achieved performances close to A1000 despite being trained on five times less

361 images. For OTU261, even the A20 and A50 classifiers achieved sensitivity and

362 precision greater than 0.86, and differences between the A20, A50 and A100

363 classifiers were not statistically significant (Figure 5).

16

364   Sensitivity in the C1000 classifier was 0.92 and 0.89 for OTU261 and OTU339,

365   respectively, which is significantly lower than A1000 (p-value <0.0001 for both –

366   Figure A2 and A3) but only a marginal difference (0.03 each). For OTU261, the C200

367   classifier achieved lower sensitivity than the A200 but they had equal precision. For

368   OTU339, precision is also the same in A200 and all C classifiers (Figure A4). Note

369   that for both OTUs, precision of all treatments in C were either not significantly or

370   barely significantly different (p-value above 0.01). Thus, C classifiers (with 52

371   classes) achieve performances almost as good as A classifiers when training on 200

372   or less images.

373   Group B classifiers tended to show slightly lower sensitivity than A classifiers and

374   slightly lower precision than C, although often not significantly different.

375

Discussion

377   In this study, our purpose was to test the capacity of a transferred CNN classifier

378   (partially trained on a different dataset) to identify benthic animals and, by extension,

379   to test if this methodology can be successfully applied in ecology by non-specialists

380   with a relatively small data set, open-source software and libraries, as well as a short

381   investment in time after manual image annotation.

Overall performances

383   Our classifiers achieved maximum average performance of 0.78 in sensitivity and

384   0.75 in precision. In other studies, performances achieved through manual

385   annotation range from 50 to 95% for benthic fauna (Beijbom et al. 2015, Durden et

386   al. 2016) and 84 to 94% accuracy for plankton (Culverhouse et al. 2003). There is no

17

387 consensus on what is an acceptable error rate in the ecological literature but, to be

388 competitive with experts, automated identification performances should be towards

389 the higher end of those achieved manually. In this regard, Culverhouse et al. (2014)

390 report an anecdotal value of 0.9 cited by experts. Previous studies on marine

391 ecosystems sampled via images that have attempted to automatically classify

392 multiple benthic megafaunal taxa with various methods sometimes achieve

393 performances comparable to those of experts. Beijbom et al. (2012) found average

394 accuracies up to 97% when classifying different coral species in shallow reefs.

395 Schoening et al. (2012) found an average sensitivity of 0.87 and precision of 67%

396 when classifying deep benthic megafauna in the Arctic. Marburg and Bigham (2016)

397 found 89% accuracy when classifying benthic mobile megafauna off the Oregon

398 coast. When considering other faunal groups, CV can achieve even higher

399 performances, for example, Siddiqui et al. (2018) classified fish species with up to

400 96.7% average accuracy.

401 Even at their best performances, our classifiers would misclassify more than one out

402 of 5 observations if they were used to make novel predictions. This is not good

403 enough to be considered a suitable replacement for manual annotation. To be the

404 tool benthic ecologists need, average performances need to be increased by at least

405 10 or 15%.

406 <u>Impact of the number of images in training on performances</u>

407 In our study, average performance measured as both sensitivity and precision

408 increased with the number of images used in training. Performances obtained with

409 1000 training images are significantly better than that obtained with fewer images but

410 only marginally so than those obtained with five times less (200) images.

411   Extrapolation of the data suggests that performances may never greatly exceed

412   those obtained with 1000 training images regardless of how many images are used.

413   It has been generally demonstrated that more data is preferable when modelling

414   (Enric et al. 2013) and training classifiers (Lu & Weng 2007, Maxwell et al. 2018).

415   Unsurprisingly then, our results suggest that the number of training images has a

416   clear positive effect on performance, particularly on sensitivity. Sun et al. (2017)

417   tested their generalist object classifiers with 10, 30 and 100 million images and

418   observed a clear increase in performance. Siddiqui et al. (2018) also found that

419   increasing the size of a dataset by 25% (20000 to 25000 images) resulted in a 6.6%

420   increase in the accuracy of the same CNN.

421   More data, however, is not a simple solution to low performances as the relationship

422   between the amount of training data and performance is not linear. Sun et al. (2017)

423   report a logarithmic relationship between the size of the training set and

424   performance. These authors gained less than 20% increase in performance by

425   adding 90 million images to their training set. This logarithmic relationship has also

426   been reported by Favret & Sieracki (2016) in their fly species classifiers. These

427   authors note a diminishing return of adding more training data and observed little

428   gain when doubling their training size from 50 to 100 images. Cho et al. (2015), who

429   classified computed tomography images of six human body parts, found the same

430   logarithmic relationship and, although it was 95.7% with 200 training images, their

431   desired 99.5% accuracy target was only reached with 4092 images. Thus, there is

432   an optimal size to every dataset and a point beyond which more training data results

433   in very little gain. This point can be determined by the goal of the study and what is

434   considered acceptable performance. With our methodology, this point occurs at 200

19

435   images, and represents a reasonable amount of manual work for ecologists aiming

436   to build the dataset to train a CNN.

437   Impact of the number of OTU classes in training on performances

438   We observed that classifiers with a small (7) number of classes had better

439   performances than those trained with 27 or 52. The difference in performance

440   between the latter two was marginal, although significant.

441   The number of classes in machine learning studies is usually driven by the dataset

442   and the research question rather than maximizing performance by limiting the

443   number of classes. Thus, few studies have assessed the effect of that number on

444   their performance. Accuracies in the 24 CV-based animal identification studies cited

445   by Favret and Sieracki (2016) and Weinstein (2018) were not significantly correlated

446   to the number of classes used in each classifier. In their large dataset experiment,

447   Sun et al. (2017) also found no difference when training with 1000 or 18000 classes.

448   But in contrast, Favret and Sieracki (2016) observed a counterintuitive increase in

449   performance as more insect species were included into their training set. They

450   hypothesised that, although a higher number of possible outcomes could increase

451   confusion, the higher number of comparison points helped determine the important

452   features of each category. Further tests are needed to disentangle the effect of the

453   number of classes in training or the relative difference in morphology of these

454   classes on performance. In general, practical applications of CV in ecology would

455   benefit from more information on this effect.

Potential application of CV to a real ecological dataset

457 To deploy classifiers such as these in a "real-life" ecological study, reasonable

458 performances must be achieved while retaining time and cost effectiveness of

459 building the training set.

460 In our study, no classifier achieved average performance above 0.78, which would

461 mean one misidentification out of 5 predictions, at best. We also observed high

462 interclass variability as some OTU were consistently well identified while others

463 were, on the contrary, always misclassified. Even if the measured average

464 performances were considered acceptable, it would introduce completely false

465 appreciation of the distribution of some OTUs and local diversity.

466 This variability in both expert and machine classification performance between

467 classes or taxa has been observed by other authors (Beijbom et al. 2015, Cho et al.

468 2015). Experts in Durden et al. (2016) had various annotation successes for different

469 taxa and Schoening et al. (2012) found that human observers and their semi-

470 automated classifier had variable success at detecting and identifying different taxa

471 but agreed on which one had the best performance. It is therefore sensible to

472 consider the predictions of each OTU class separately and only rely on those for

473 which the classifier achieves good performances, both in precision and sensitivity.

474 Good performance obtained by our classifier with some specific OTU classes is

475 encouraging and automated annotations could be an appropriate method to study

476 them. The top 10 best and worst OTUs ranked by sensitivity shows that the

477 classifiers are better at identifying large sized organisms exhibiting a low intra-class

478 morphological variability. The majority of the top 10 OTUs were rare (e.g. less than

479 100 training images). If CV were applied to these rare taxa, there would be a

480   proportionally higher impact of any misidentification or false positive results. Given

481   their relatively low number of occurrences (tens to a few hundreds), a manual

482   verification step (or semi-automated identification), as performed by Schoening et al.

483   (2012) and suggested by Marburg and Bigham (2016), would be easy to perform for

484   a reasonable time investment and ensure the reliability of the predictions. On the

485   other hand, OTU261 and OTU339, both among the top 10 OTU classes, were very

486   abundant in the study area (above 1200 individuals). As manual validation of

487   identification of these OTUs would be impractical, their identification should be fully

488   automated if the classifier is to be deployed on a larger dataset.

489   With OTU261 and OTU339, high sensitivity (up to 0.95 and 0.92 respectively) and

490   high precision (up to 0.95 and 0.82, respectively) were achieved by the classifiers,

491   meaning they were usually correctly identified and false positives were relatively

492   rare. These performances are equivalent to those of human experts working on a

493   very similar ecosystem (Durden et al. 2016) without the inconsistency over time by

494   individual observers reported by these authors. Therefore, these classifiers could be

495   applied to remaining un-annotated images in our dataset and provide useful

496   presence records of these specific OTUs. This would be a valuable contribution to

497   this study of deep-sea ecosystems.

498   Classifier A1000 had the best performance of all classifiers and would detect almost

499   all individuals of OTU261 and OTU339, but it needs a large training set, while the

500   A200 classifier has very similar performances but needs five time less training

501   material and is therefore more cost-effective. These group A classifiers however, risk

502   producing a high amount of false positives if they encounter too many individuals of

503   an OTUs they have not been trained on. Thus, it is only applicable if diversity at the

504   study site is low or it is predominantly represented by a small number of OTUs.

22

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

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

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

526 Even if the presence records of some OTUs are not sufficient to understand the
527 composition and dynamics of an ecosystem, it will still contribute to it and more
528 importantly, it will take-on some of the annotation time, leaving experts free to

23

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.

24

554

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.

574

Acknowledgments

25

582

References

584   Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A,
585   Dean J, Devin M (2016) Tensorflow: Large-scale machine learning on
586   heterogeneous distributed systems. arXiv preprint arXiv:160304467
587   Althaus F, Hill N, Ferrari R, Edwards L, Przeslawski R, Schönberg CH, Stuart-Smith
588   R, Barrett N, Edgar G, Colquhoun J (2015) A standardised vocabulary for
589   identifying benthic biota and substrata from underwater imagery: the CATAMI
590   classification scheme. PloS one 10:e0141039
591   Assembly UNG (2003) Oceans and the Law of the Sea. Report of the Secretary
592   General. A/58/65
593   Ayma A, Aguzzi J, Canals M, Lastras G, Bahamon N, Mecho A, Company JB (2016)
594   Comparison between ROV video and Agassiz trawl methods for sampling
595   deep water fauna of submarine canyons in the Northwestern Mediterranean
596   Sea with observations on behavioural reactions of target species. Deep Sea
597   Research Part I: Oceanographic Research Papers 114:149-159
598   Beijbom O, Edmunds PJ, Kline DI, Mitchell BG, Kriegman D Automated annotation
599   of coral reef survey images. Proc Computer Vision and Pattern Recognition
600   (CVPR), 2012 IEEE Conference on. IEEE
601   Beijbom O, Edmunds PJ, Roelfsema C, Smith J, Kline DI, Neal BP, Dunlap MJ,
602   Moriarty V, Fan T-Y, Tan C-J (2015) Towards automated annotation of
603   benthic survey images: Variability of human experts and operational modes of
604   automation. PloS one 10:e0130312
605   Bicknell AW, Godley BJ, Sheehan EV, Votier SC, Witt MJ (2016) Camera technology
606   for monitoring marine biodiversity and human impact. Frontiers in Ecology and
607   the Environment 14:424-432
608   Borja A, Elliott M, Snelgrove PVR, Austen MC, Berg T, Cochrane S, Carstensen J,
609   Danovaro R, Greenstreet S, Heiskanen A-S, Lynam CP, Mea M, Newton A,
610   Patrício J, Uusitalo L, Uyarra MC, Wilson C (2016) Bridging the Gap between
611   Policy and Science in Assessing the Health Status of Marine Ecosystems.
612   Frontiers in Marine Science 3
613   Brandt A, Gutt J, Hildebrandt M, Pawlowski J, Schwendner J, Soltwedel T, Thomsen
614   L (2016) Cutting the Umbilical: New Technological Perspectives in Benthic
615   Deep-Sea Research. Journal of Marine Science and Engineering 4:36
616   Bullimore RD, Foster NL, Howell KL (2013) Coral-characterized benthic
617   assemblages of the deep Northeast Atlantic: defining "Coral Gardens" to
618   support future habitat mapping efforts. ICES J Mar Sci 70:511-522

26

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

28

719 Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, Clune
720   J (2018) Automatically identifying, counting, and describing wild animals in
721   camera-trap images with deep learning. Proceedings of the National Academy
722   of Sciences:201719367
723 Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Transactions on
724   knowledge and data engineering 22:1345-1359
725 Pimentel MAF, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty
726   detection. Signal Processing 99:215-249
727 Rampasek L, Goldenberg A (2016) TensorFlow: Biology's Gateway to Deep
728   Learning? Cell Syst 2:12-14
729 Rex MA, Etter RJ (2010) Deep-sea biodiversity: pattern and scale. Harvard
730   University Press
731 Ripley B, Venables W, Ripley MB (2016) Package 'nnet'. R package version:7-3
732 Roberts JM, Harvey SM, Lamont PA, Gage JD, Humphery JD (2000) Seabed
733   photography, environmental assessment and evidence for deep-water
734   trawling on the continental margin west of the Hebrides. Hydrobiologia
735   441:173-183
736 Rohlf FJ, Sokal RR (1967) Taxonomic structure from randomly and systematically
737   scanned biological images. Systematic Zoology 16:246-260
738 Romero-Ramirez A, Grémare A, Bernard G, Pascal L, Maire O, Duchêne JC (2016)
739   Development and validation of a video analysis software for marine benthic
740   applications. Journal of Marine Systems 162:4-17
741 Rowden AA, Schnabel KE, Schlacher TA, Macpherson E, Ahyong ST, de Forges BR
742   (2010) Squat lobster assemblages on seamounts differ from some, but not all,
743   deep-sea habitats of comparable depth. Marine Ecology-an Evolutionary
744   Perspective 31:63-83
745 Schneider S, Taylor GW, Kremer SC (2018) Deep Learning Object Detection
746   Methods for Ecological Camera Trap Data. arXiv preprint arXiv:180310842
747 Schoening T, Bergmann M, Ontrup J, Taylor J, Dannheim J, Gutt J, Purser A,
748   Nattkemper TW (2012) Semi-automated image analysis for the assessment of
749   megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN.
750   PloS one 7:e38179
751 Schoening T, Durden J, Preuss I, Albu AB, Purser A, De Smet B, Dominguez-Carrió
752   C, Yesson C, de Jonge D, Lindsay D (2017) Report on the Marine Imaging
753   Workshop 2017. Research Ideas and Outcomes 3:e13820
754 Siddiqui SA, Salman A, Malik MI, Shafait F, Mian A, Shortis MR, Harvey ES,
755   Handling editor: Howard B (2018) Automatic fish species classification in
756   underwater videos: exploiting pre-trained deep neural network models to
757   compensate for limited labelled data. ICES J Mar Sci 75:374-389
758 Solan M, Germano JD, Rhoads DC, Smith C, Michaud E, Parry D, Wenzhöfer F,
759   Kennedy B, Henriques C, Battle E, Carey D, Iocco L, Valente R, Watson J,
760   Rosenberg R (2003) Towards a greater understanding of pattern, scale and
761   process in marine benthic systems: a picture is worth a thousand worms.
762   Journal of Experimental Marine Biology and Ecology 285-286:313-338
763 Sorensen S, Treible W, Hsu L, Wang X, Mahoney AR, Zitterbart DP, Kambhamettu
764   C Deep Learning for Polar Bear Detection. Proc Scandinavian Conference on
765   Image Analysis. Springer
766 Sun C, Shrivastava A, Singh S, Gupta A Revisiting Unreasonable Effectiveness of
767   Data in Deep Learning Era. Proc 2017 IEEE International Conference on
768   Computer Vision (ICCV)

29

769 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z Rethinking the inception
770   architecture for computer vision. Proc Proceedings of the IEEE Conference on
771   Computer Vision and Pattern Recognition
772 Team RC (2014) R: A Language and Environment for Statistical Computing. Vienna;
773   2014.
774 Van Dover C, Aronson J, Pendleton L, Smith S, Arnaud-Haond S, Moreno-Mateos D,
775   Barbier E, Billett D, Bowers K, Danovaro R (2014) Ecological restoration in
776   the deep sea: Desiderata. Marine Policy 44:98-106
777 Wei C-L, Rowe GT, Escobar-Briones E, Boetius A, Soltwedel T, Caley MJ, Soliman
778   Y, Huettmann F, Qu F, Yu Z (2010) Global patterns and predictions of
779   seafloor biomass using random forests. PLoS One 5:e15323
780 Weinstein BG (2018) A computer vision for animal ecology. Journal of Animal
781   Ecology 87:533-545
782 Wheeler B, Torchiano M (2010) lmPerm: Permutation tests for linear models. R
783   package version 1
784 Wickham H (2017) Tidyverse: Easily install and load'tidyverse'packages. R package
785   version 1
786 Williams SB, Pizarro O, Steinberg DM, Friedman A, Bryson M (2016) Reflections on
787   a decade of autonomous underwater vehicles operations for marine survey at
788   the Australian Centre for Field Robotics. Annual Reviews in Control 42:158-
789   165
790 Wynn R, Bett B, Evans A, Griffiths G, Huvenne V, Jones A, Palmer M, Dove D,
791   Howe J, Boyd T (2012) Investigating the feasibility of utilizing AUV and Glider
792   technology for mapping and monitoring of the UK MPA network.
793   Southampton: National Oceanography Centre

794

Tables

796 *Table 1: Nomenclature of classifiers names and characteristics. The different classifiers names are a combination*
797 *of group name and image numbers per Operational Taxonomical Units (OTU) in training. Groups are defined by*
798 *the number of different OTUs (or classes) in the training set. In the different groups, the OTUs used are those for*
799 *which the minimum number of images indicated are available. Within each group, treatments refer to the number*
800 *of images of each class in training. The same treatments (20, 50, 100, 200, 500 and 1000 images per OTU in*
801 *training) were applied to each group but only the classifiers names **in bold** are balanced (equal number of*
802 *images for every class). In unbalanced designs, the maximum number of available images is used and is*
803 *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

804

30

805

806

807

808
809 *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.*

810

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

31

811

Figures

813
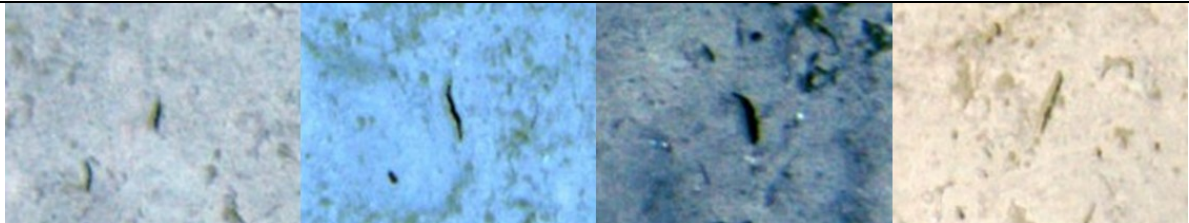
OTU603: Very small elongated sponge. Shape is constant.
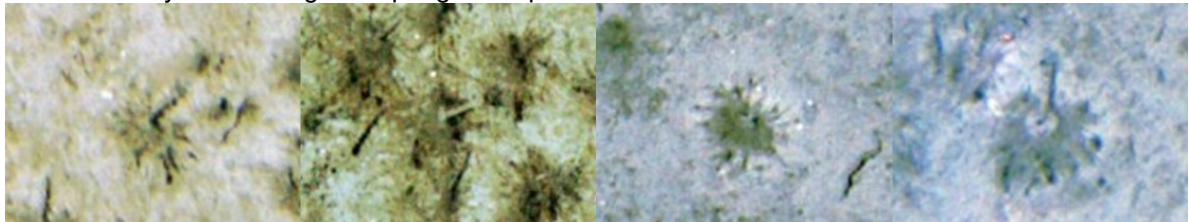
OTU375: Small tube worm. The gills can hide the tube
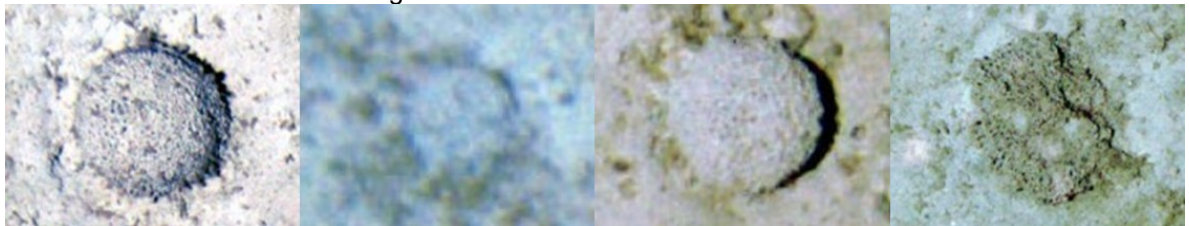
OTU261: The xenophyophores *Syringamina fragillissima*
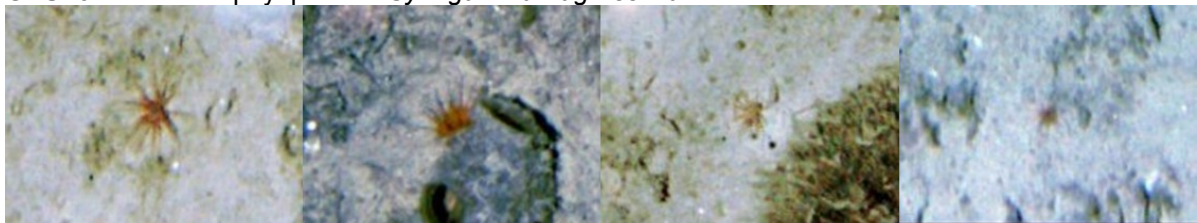
OTU23: Small halcampid/edwardsiid anemone
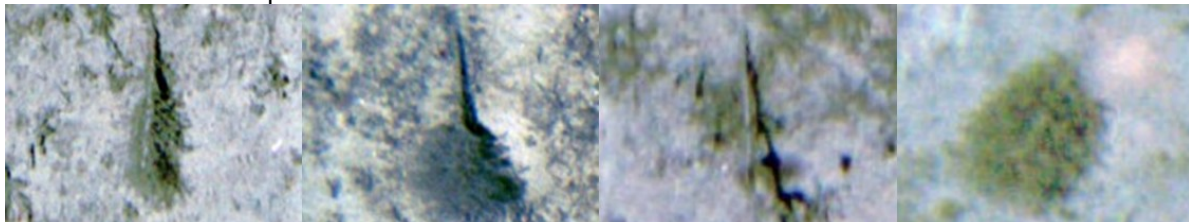
OTU995: Unknown animal, possibly a chrisogorgid

OTU2: Cerianthid anemone of various size
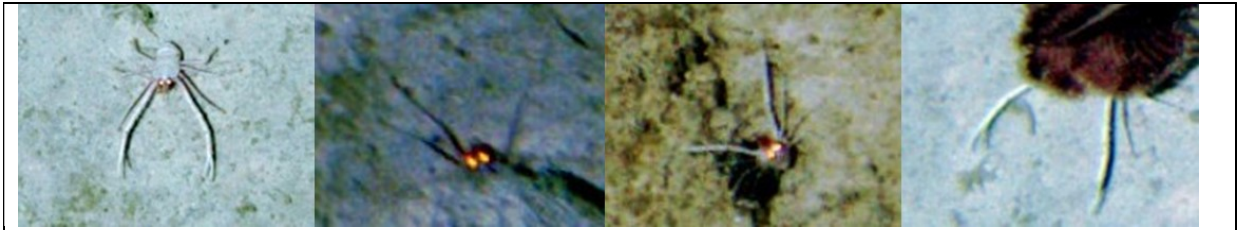
OTU339: The squat lobster *Munida sarsi/tenuimana*

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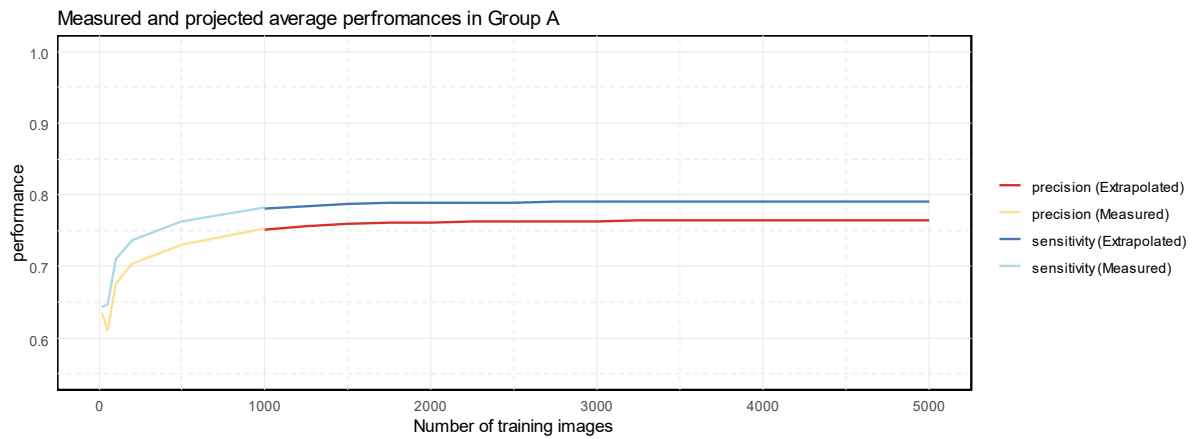
Figure 2: Classifier performances (sensitivity and precision) per number of training images measured (20 – 1000) and extrapolated (1000 – 10000). Grey dots show averaged values across all OTUs for each classifiers

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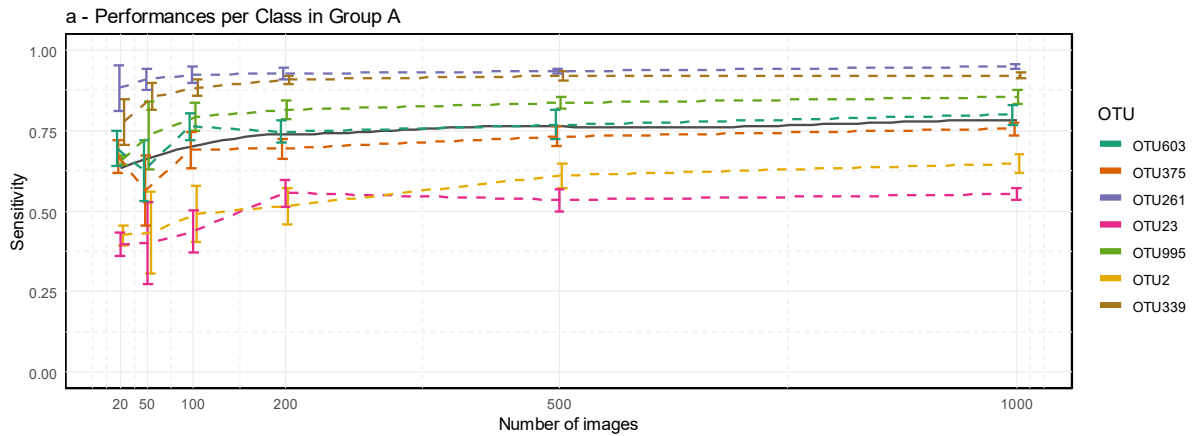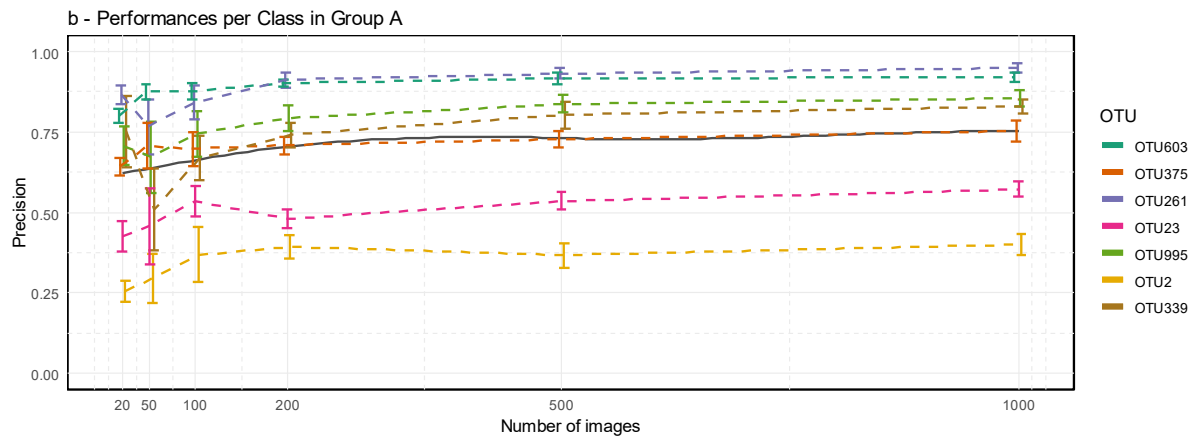
Figure 3: a) Evolution of Sensitivity in Group A classifier trained with an increasing number of images. b) Differences in Precision in Group A classifier trained with an increasing number of images. The black line is 'loess' smoothed curve of the average of all the classes and greyed area is a t-based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

a - Sensitivity per number of classes in training

b - Precision per number of classes in training
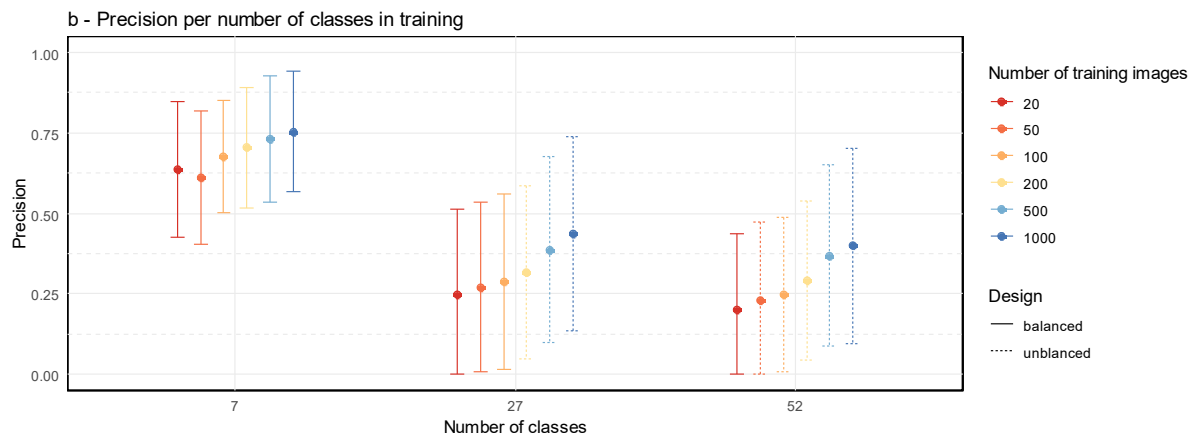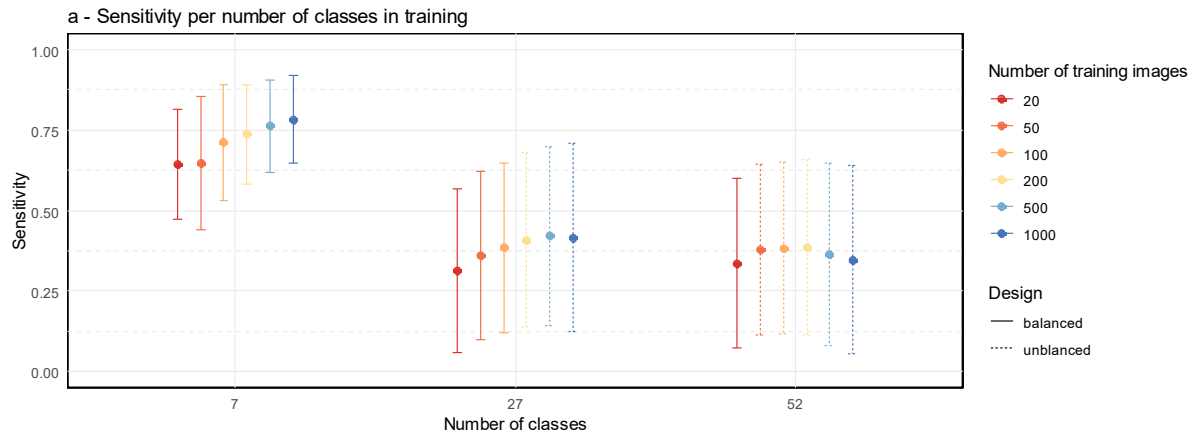
Figure 4: a) Differences in sensitivity in classifiers trained with different number of classes and images. b) Differences in precision in classifier trained with different number of classes and images. Error bars are standard deviation of the 10 random splits.
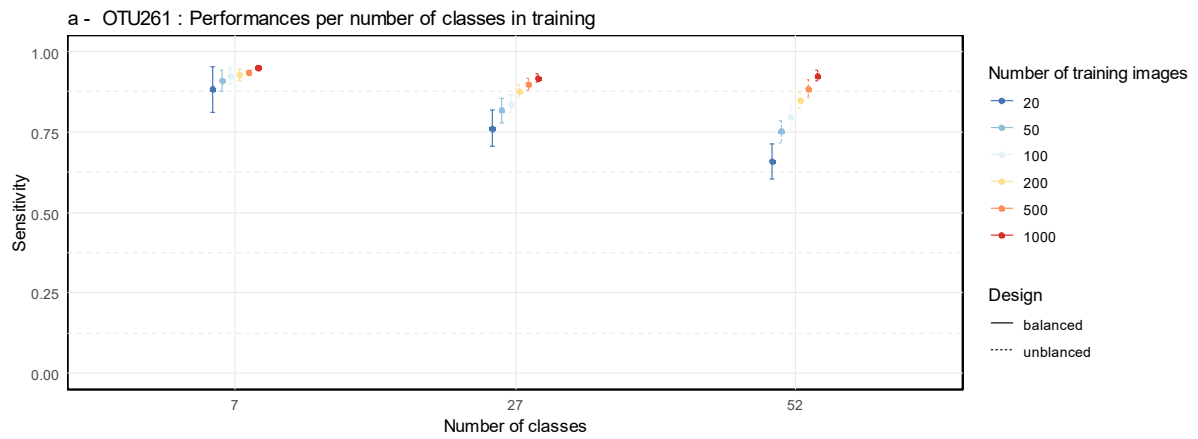
832

833
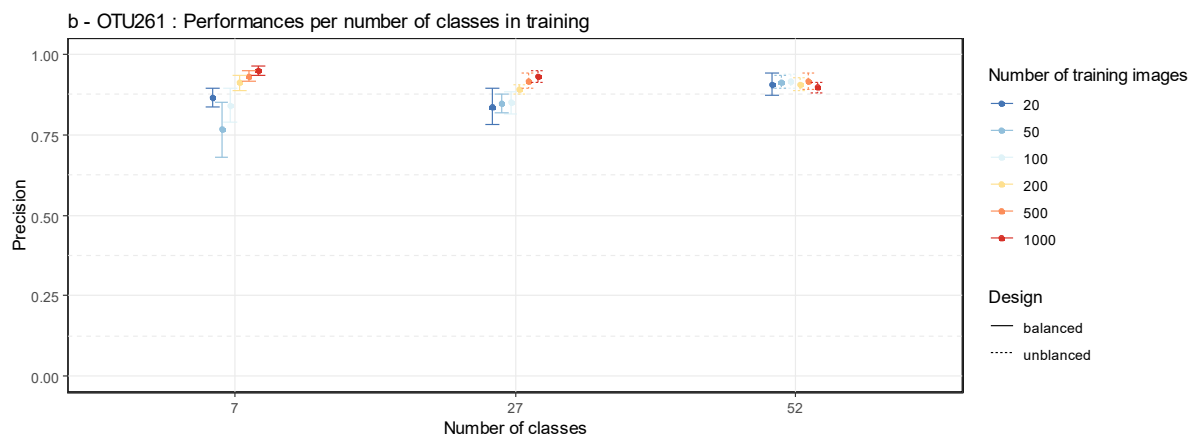
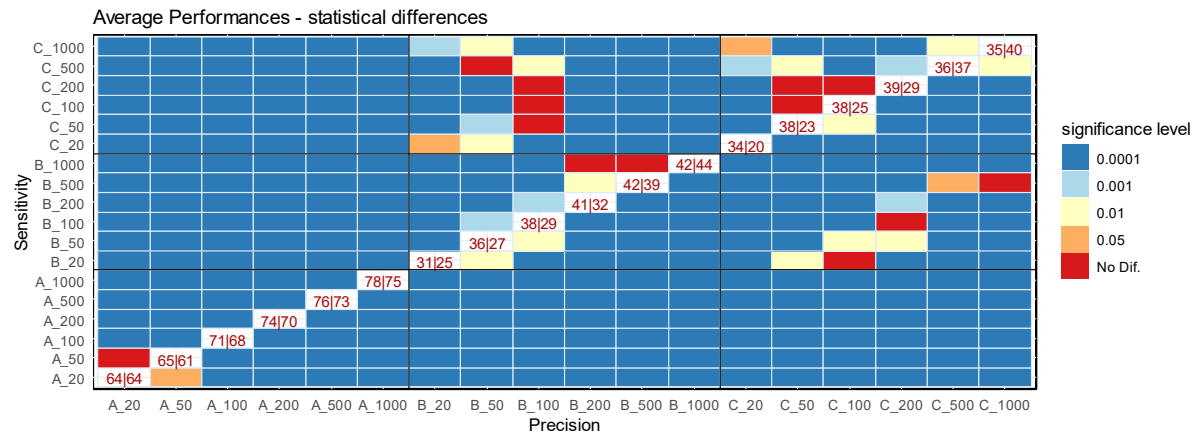834 *Figure 5: a) Differences in sensitivity for OTU261 in classifier trained with different number of classes and*
835 *images. Error bars are standard deviation calculated from the 10 random splits. b) Differences in precision for*
836 *OTU261 in classifier trained with different number of classes and images. Error bars are standard deviation of*
837 *the 10 random splits*

838

839

840

841

842

843

Appendix

846

847

848 *Figure A1: Pair-wise permutation-based analysis of variance of differences in sensitivity (upper left) and precision*
849 *(lower right) between each treatment. The numbers in central cells indicates sensitivity (left) and precision (right)*
850 *of corresponding treatments on the axis. Significance level indicate at which alpha threshold the two treatments*
851 *are significantly different in percentages of maximal value (i.e. 1). No dif. indicates a p-value above 0.05.*
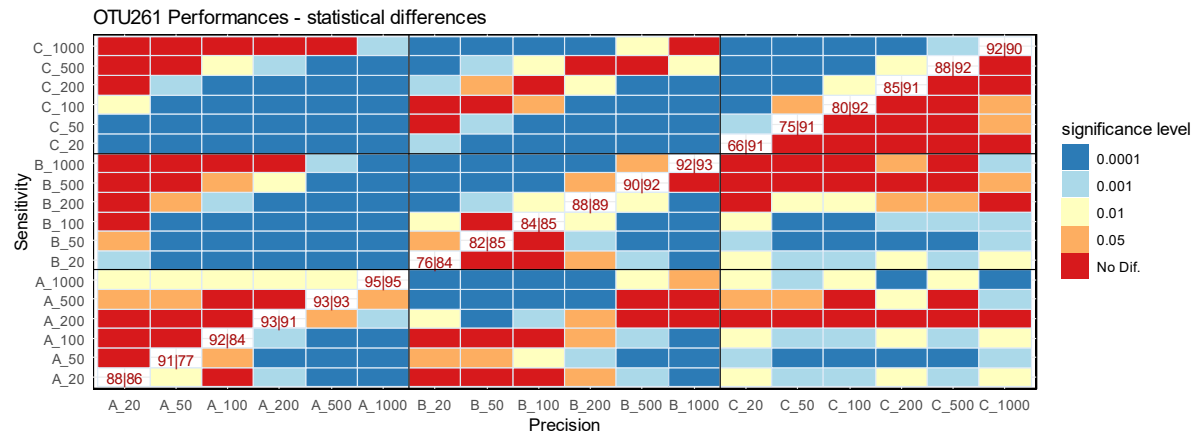
852

853

854 *Figure A2: Pair-wise permutation-based analysis of variance of differences in sensitivity (upper left) and precision*
855 *(lower right) between each treatment. The numbers in central cells indicates sensitivity (left) and precision (right)*
856 *of corresponding treatments on the axis in percentages of maximal value (i.e. 1). Significance level indicate at*
857 *which alpha threshold the two treatments are significantly different. No dif. indicates a p-value above 0.05.*
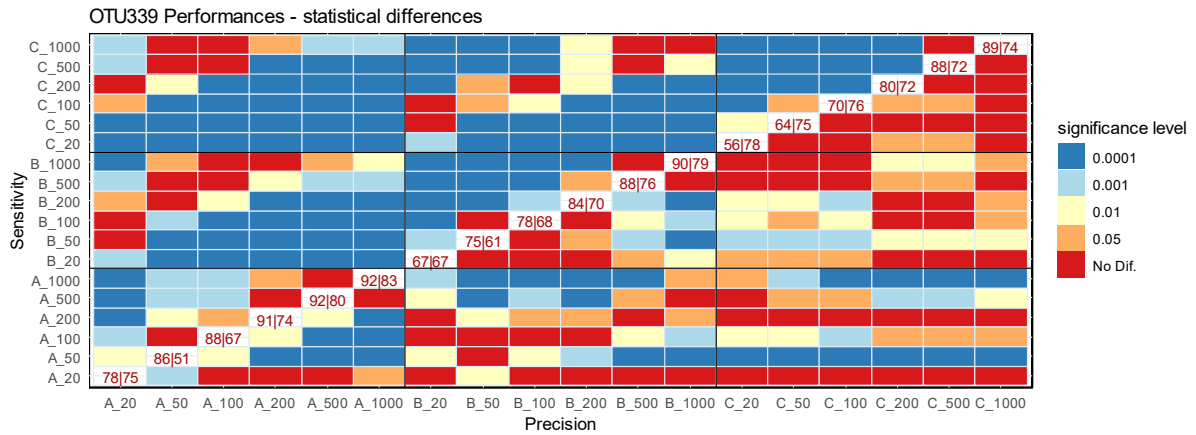
858

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.
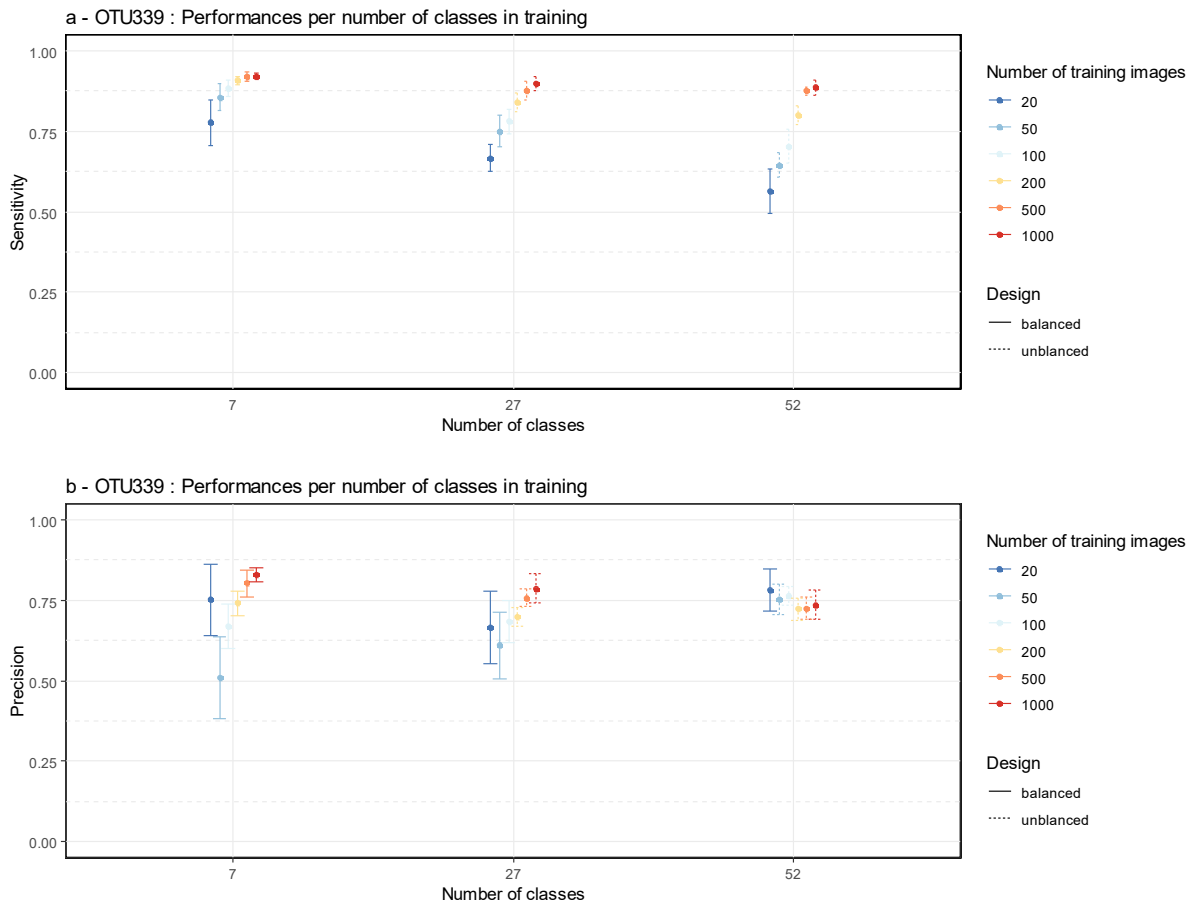
859
860
861
862

863

864

865

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.

866
867
868
869