SCIENTIFIC REPORTS
natureresearch

OPEN

Annotation-free learning of plankton for classification and anomaly detection

Vito P. Pastore[1,2]✉, Thomas G. Zimmerman[1,2], Sujoy K. Biswas[1,2] & Simone Bianco[1,2]

The acquisition of increasingly large plankton digital image datasets requires automatic methods of recognition and classification. As data size and collection speed increases, manual annotation and database representation are often bottlenecks for utilization of machine learning algorithms for taxonomic classification of plankton species in field studies. In this paper we present a novel set of algorithms to perform accurate detection and classification of plankton species with minimal supervision. Our algorithms approach the performance of existing supervised machine learning algorithms when tested on a plankton dataset generated from a custom-built lensless digital device. Similar results are obtained on a larger image dataset obtained from the Woods Hole Oceanographic Institution. Additionally, we introduce a new algorithm to perform anomaly detection on unclassified samples. Here an anomaly is defined as a significant deviation from the established classification. Our algorithms are designed to provide a new way to monitor the environment with a class of rapid online intelligent detectors.

Plankton are a class of aquatic microorganisms, composed of both drifters and swimmers, which can vary significantly in size, morphology, and behavior. The exact number of plankton species is not known, but an estimation of oceanic plankton puts the number between 3,444 and 4,375[1]. Plankton are at the bottom of the aquatic food chain and marine phytoplankton are estimated to be responsible for over 50% of all global primary production[2] and play a fundamental role in climate regulation. Thus, changes in plankton ecology may have a profound impact on global climate, as well as deep social and economic consequences. It seems therefore paramount to collect and analyze real time plankton data to understand the relationship between the health of plankton and the health of the environment they live in. Traditionally, plankton are surveyed using either satellite remote sensing, where leftover biomass is inferred indirectly through measurement of total chlorophyll concentration, or with large net tows via oceanic vessels[3], with subsequent microscopic analysis of the preserved samples. Satellite imaging methods are extremely accurate in terms of global geographic association and very useful for broad species characterization but may present practical challenges in terms of accuracy of the performed counts, species preservation and fine-grained characterization. The analysis of preserved samples, instead, allows for fine grained classification and accurate counting with narrow spatial sampling. More recently, real time observation of plankton species has been made possible by novel instruments for high-throughput in situ autonomous and semi-autonomous microscopy[4]. Such high-resolution imaging instruments make it possible to observe and study spatio-temporal changes in plankton morphology and behavior, which can be correlated with environmental perturbations. Sudden or unexpected changes in number, shape, aggregation patterns, population composition or collective behavior may be used to infer anomalous conditions related to potentially catastrophic events, either natural, like harmful algal blooms, or man-made, like industrial run offs or oil spills. Intelligent systems trained on curated data could help establish the characteristics of a healthy ecosystem and detect perturbations that may represent potential threats. More importantly, given the diversity of plankton morphology and behavior across species and the growing but still limited availability of high-quality labeled data sources, there is a need for algorithms which require minimal supervision to classify and monitor plankton species with a performance approaching that of supervised algorithms. Moreover, it is also desirable for such algorithms to aid the discovery of new plankton classes, which cannot generally happen with supervised classification techniques. In this paper we propose a set of novel algorithms to reliably characterize and classify plankton data. Our method is based

[1]Industrial and Applied Genomics, AI and Cognitive Software, IBM Research – Almaden, San Jose, CA, USA. [2]NSF Center for Cellular Construction, University of California San Francisco, San Francisco, CA, USA. ✉email: vitopaolopastore@gmail.com

on an unsupervised approach to overcome the limits of supervised machine learning techniques and designed to dynamically classify plankton from instruments that continuously acquire plankton images. First, we evaluate the performances of our algorithms on a mixture of ten freshwater plankton species imaged with a lensless microscope designed for in situ data collection[5]. Next, we evaluate the performance of our algorithms on an image dataset extracted from the Woods Hole Oceanographic Institution (WHOI) plankton database[6]. Machine learning methods are becoming a popular way to characterize and classify plankton[7–14].

In[15], the authors developed an automated analysis system for the identification of phytoplankton using neural networks. A recent paper[16] explores the use of Convolutional Neural Networks to classify species of zooplankton, by introducing an architecture named ZooplanktoNet. The authors claim that their customized architecture can reach higher accuracy compared to standard deep learning configurations, like VGG, AlexNet, CaffeNet, and GoogleNet. In[17] and [18], the authors use an SVM based algorithm to classify species with high accuracy from the WHOI dataset. In a recent Kaggle competition contest (https://www.kaggle.com/c/datasciencebowl), the authors developed a deep learning architecture named DeepSea[19] to perform accurate classification of plankton collected with an underwater camera. In[20] the authors combine features obtained with multiple kernel learning to achieve higher accuracy than classic machine learning algorithms. However, all these advancements use supervised learning algorithms that rely on large labeled training sets which are very difficult and time consuming to create. Although recent computational advances may reduce the annotation burden for large biological datasets[21], a high-performance unsupervised learning algorithm can provide an alternative for real time unbiased in situ analysis.

Results

Plankton classifier.

We developed an unsupervised customized pipeline for plankton classification and anomaly detection, that we named plankton classifier. The pipeline, shown in Fig. 1, is tested on a collection of videos containing ten freshwater species of plankton captured with a lensless microscope[5]. Each video is ten seconds long and contains one or more species. As the method is unsupervised, no labels are provided to the classifier during training. The plankton classifier consists of four modules: an image processor, a feature extractor, an unsupervised partitioning module, and a classification module. The image processor examines each frame of video and generates cropped images of each plankter. The feature extractor examines each plankter image and generates a collection of features. The unsupervised partitioning module clusters samples by features into classes. The classification module comprises of a neural network-based anomaly detector to both perform classification based on the inferred labels and provide information to extend the database in an unsupervised manner. A sample is considered an anomaly with respect to a class if the extracted features are significantly different from the class average, as described below. The classification module also includes a standard neural network classifier, for performance comparison. See section materials and methods for a description of the modules in more details, along with the methods considered and tested that led to our final design.

Unsupervised partitioning performance.

First, the plankton classifier examines each frame of an acquired video and generates cropped images of each plankter A set of 131 features is then extracted, as described in Materials and Methods. The unsupervised partitioning module uses such features to place each plankton sample into one of Z classes. To automatically obtain the number of classes Z from the dataset, we have designed a custom algorithm based on Partition Entropy (PE). To ensure high confidence for the estimation, we perform ten iterations taking the mode of the resulting estimated number of clusters distribution as the predicted Z (see Materials and Methods). We evaluated the robustness of the implemented method on random subsets of the lensless dataset with different sizes, ranging from three to ten species. The box plot indicating the distribution for the estimated number of clusters Z among ten iterations is shown in Fig. 2e. The inferred number of classes, Z, is identified with high confidence in every case. A comparison of the performance of this algorithm against other existing methods is reported in the Supporting Information. Once we have obtained the number of clusters, we compared three clustering algorithms (see Supporting Information): k-Means, Fuzzy k-Means and Gaussian Mixture Model (GMM). Clustering accuracy is evaluated using purity (see materials and methods). The Fuzzy k-Means algorithm reaches a purity value of 0.934 (see Fig. 2a, b), outperforming the standard k-Means (purity value = 0.887) and GMM[22] (purity value = 0.886). A posterior analysis of the results of the GMM reveals that this algorithm is not able to distinguish between Blepharisma americanum and Paramecium bursaria, due to their nearly identical appearance in the acquired videos. The Fuzzy k-Means algorithm is able to match the fuzziness exhibited by the plankton classes in parameter space which explains the lower accuracy of the crisp algorithms (k-Means and GMM). Therefore, we use the Fuzzy k-Means for our unsupervised classifier. A potentially important effect on the performance of any clustering algorithm is the class imbalance. The lensless microscope dataset is composed of 500 training samples for each of the ten considered species. To evaluate the impact of class imbalance, we performed the following experiment: We have built a dataset where the number of images of a species is a fraction (between 10 and 80%) of the number of images of the other species. We then evaluate the purity of this dataset and repeat the procedure for all the other species. Figure 2f reports the average performance over the ten datasets obtained as described above, as measured by the purity. The algorithm is always able to infer the correct number of species, without any overlap, with a minimum average purity value of 0.74 ± 0.09 (corresponding to 80% of class imbalance) and a maximum average purity value equal to 0.90 ± 0.08 (corresponding to 10% of class imbalance), with a maximum purity value of 0.972. This result shows that our pipeline can accurately cluster the data even in the case of strong class imbalance.

Algorithm performance on features extracted using deep feature extraction.

Feature selection is an important part of any unsupervised learning pipeline. Indeed, hand engineering features introduces a degree of arbitrariness, which can be removed using a method of automated feature selection. Deep feature

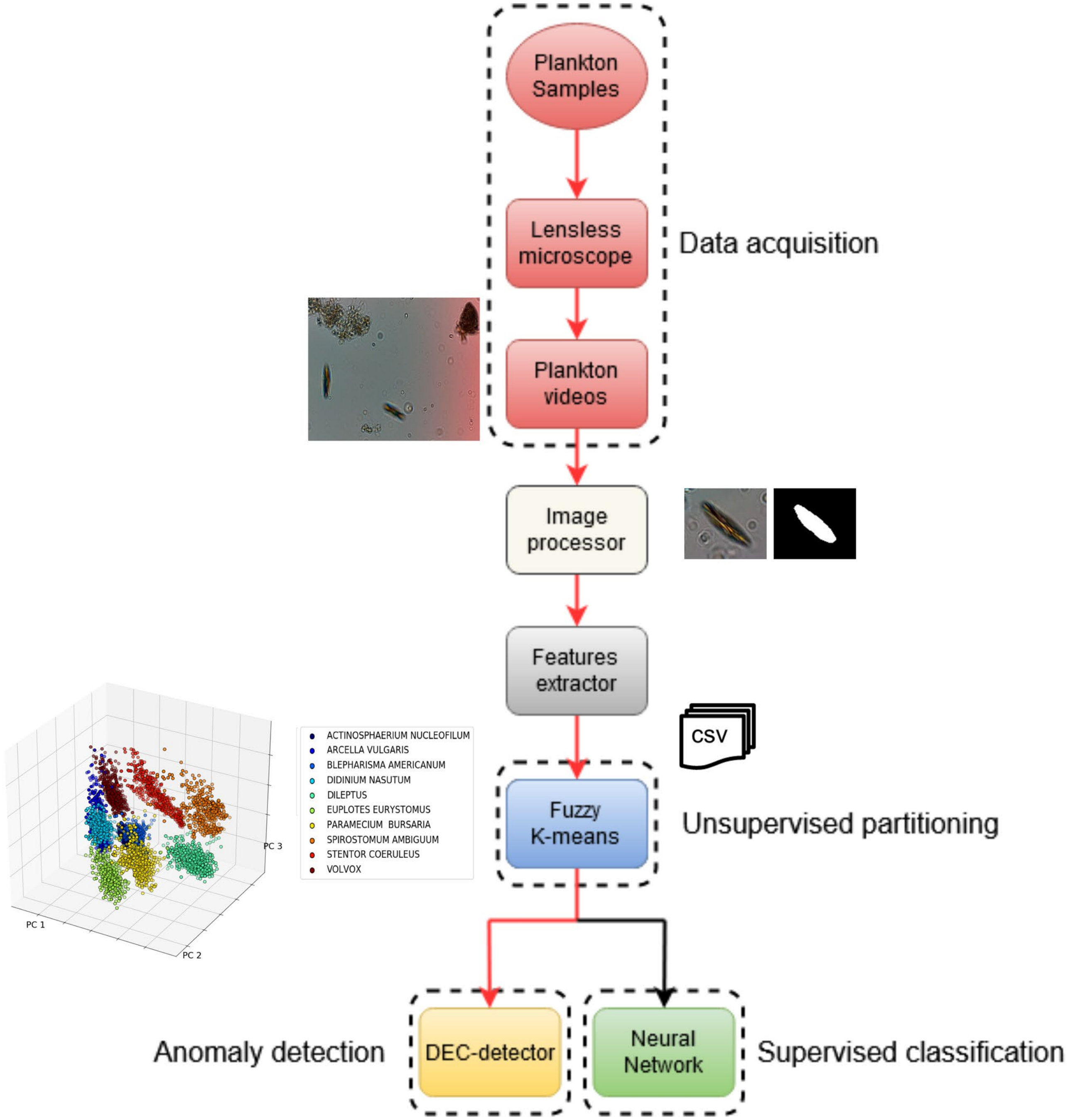
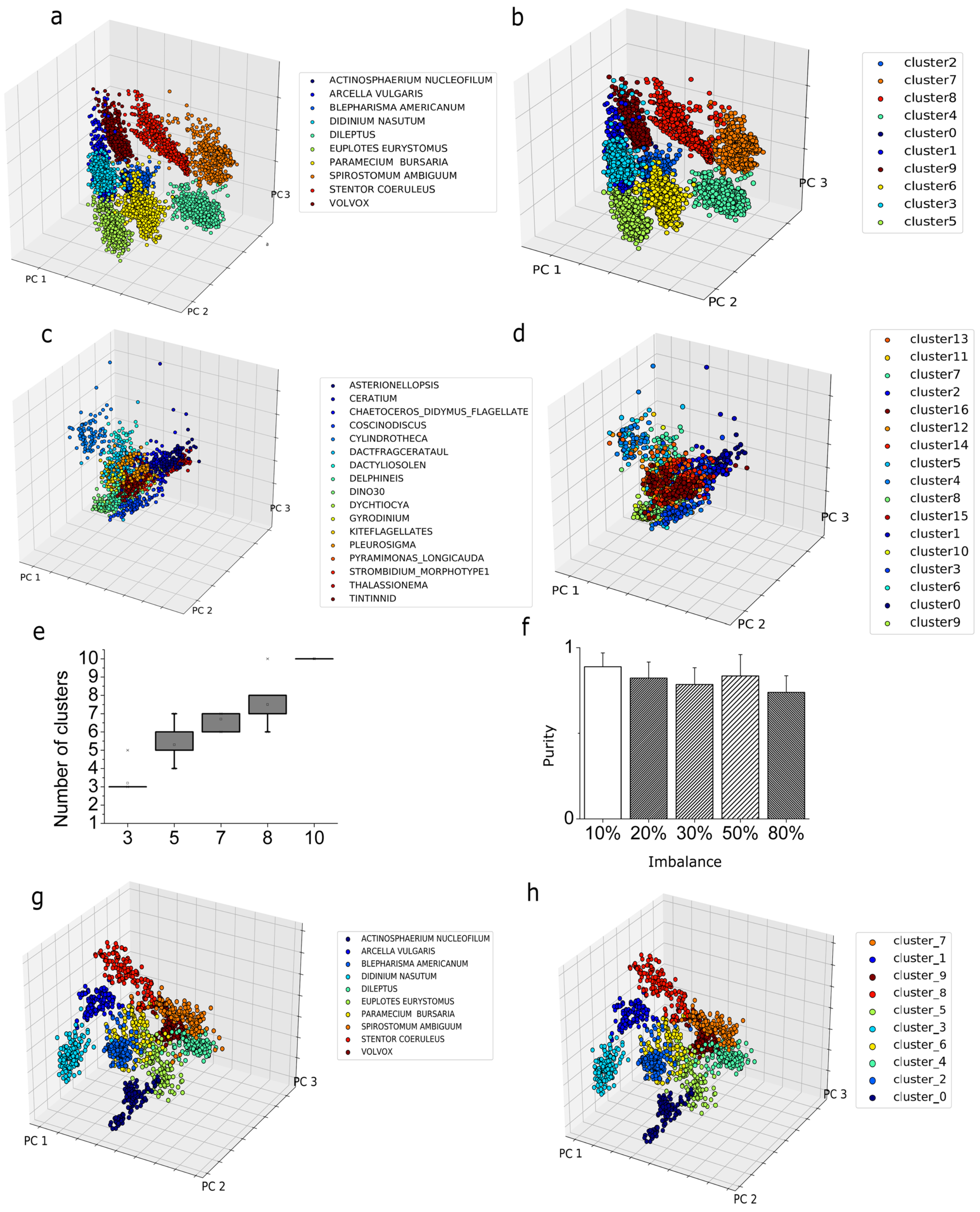


Figure 1. Schematic overview of the pipeline used to *detect and classify plankton species with minimal supervision. Our preferred embodiment is represented by the red lines.*

extraction, which consists in training a neural network architecture on either in- or out-of-domain data and use the last layer before prediction to extract features[9,23], is one such method. We trained the model described in section *Convolutional Neural Network (CNN) for deep features extraction* using the ten classes included in our lensless microscope dataset. The model reached 99% of training accuracy, 99% of validation accuracy and 98% of testing accuracy on the dataset obtained using our lensless microscope. Finally, the 128 neurons from the fully connected layers preceding the output are extracted and used as features for our pipeline. The PCA computed for the lensless microscope testing set among these features can be visualized in Fig. 2g. Figure 2h shows the results of the unsupervised partitioning procedure. The underlying structure of the data set is very accurately captured, with a purity value of 0.98. Despite the fact that the accuracy obtained using deep feature extraction is slightly higher than the one obtained using the hand engineered features (purity of 0.980 vs. 0.934), to preserve the unsupervised nature of the pipeline, we decide to use the interpretable features described in Table 1. Establishing a relationship between the morphology and the environmental perturbation can be achieved by using a subset of the hand-engineered features in Table 1 (e.g., the shape-based features or the color information), regardless of the feature extraction method. In fact, as shown in supplementary material (see Fig. S13), the whole set of hand


a
ACTINOSPHAERIUM NUCLEOFILUM
ARCELLA VULGARIS
BLEPHARISMA AMERICANUM
DIDINIUM NASUTUM
DILEPTUS
EUPLOTES EURYSTOMUS
PARAMECIUM BURSARIA
SPIROSTOMUM AMBIGUUM
STENTOR COERULEUS
VOLVOX
PC 1
PC 2
PC 3
b
cluster2
cluster7
cluster8
cluster4
cluster0
cluster1
cluster9
cluster6
cluster3
cluster5
PC 1
PC 2
PC 3
c
ASTERIONELLOPSIS
CERATIUM
CHAETOCEROS_DIDYMUS_FLAGELLATE
COSCINODISCUS
CYLINDROTHECA
DACTFRAGCERATAUL
DACTYLIOSOLEN
DELPHINEIS
DINO30
DYCHTIOCYA
GYRODINIUM
KITEFLAGELLATES
PLEUROSIGMA
PYRAMIMONAS_LONGICAUDA
STROMBIDIUM_MORPHOTYPE1
THALASSIONEMA
TINTINNID
PC 1
PC 2
PC 3
d
cluster13
cluster11
cluster7
cluster2
cluster16
cluster12
cluster14
cluster5
cluster4
cluster8
cluster15
cluster1
cluster10
cluster3
cluster6
cluster0
cluster9
PC 1
PC 2
PC 3
e
Number of clusters
10
9
8
7
6
5
4
3
2
1
3
5
7
8
10
f
Purity
1
0
10%
20%
30%
50%
80%
Imbalance
g
ACTINOSPHAERIUM NUCLEOFILUM
ARCELLA VULGARIS
BLEPHARISMA AMERICANUM
DIDINIUM NASUTUM
DILEPTUS
EUPLOTES EURYSTOMUS
PARAMECIUM BURSARIA
SPIROSTOMUM AMBIGUUM
STENTOR COERULEUS
VOLVOX
PC 1
PC 2
PC 3
h
cluster_7
cluster_1
cluster_9
cluster_8
cluster_5
cluster_3
cluster_6
cluster_4
cluster_2
cluster_0
PC 1
PC 2
PC 3

◂ **Figure 2.** Unsupervised clustering results. **a**,**b** We performed a PCA analysis on the lensless digital microscope dataset to provide a graphical representation of the data distribution into the features space. We plot the first three principal components that account for ~ 67% of the total variance. We assigned different colors to the different plankton species. a Species are assigned using ground truth labels. **b** Species are assigned to the most overlapping cluster resulting from the unsupervised partitioning procedure. **c**,**d** Same analysis and procedure applied on the WHOI dataset. **c** Species are assigned using ground truth labels. **d** Species are assigned to the most overlapping cluster, resulting from the unsupervised partitioning procedure. **e** Distribution of number of clusters computed using our PE algorithm for a random subset of species in the lensless microscope dataset. Results are reported for different initial number of species. **f** Effect of class imbalance. For each of the ten species included into the lensless microscope dataset, we simulated class imbalance by increasing the number of images available to the clustering algorithm for the considered species. **g**,**h** PCA analysis on the lensless digital microscope dataset provides a graphical representation of the data distribution into the deep features space. The unsupervised partitioning using deep features is highly accurate. The first three principal components are plotted and different colors to the different plankton species are assigned. **g** Species are assigned using ground truth labels. **h** Species are assigned to the most overlapping cluster resulting from the unsupervised partitioning.

engineered features is only needed to maximize the classification accuracy. On the other hand, for the purpose of organism classification, the customized deep feature extraction algorithm we implemented is a very viable alternative to the one proposed.

Classification.

Supervised classifier. At this stage of the pipeline, all samples have been assigned labels which have no correspondence to the actual plankton classes. We use the same trained clustering algorithm to classify the test samples, assigning each sample to the closest centroid. Using the trained Fuzzy k-means algorithm we reach a testing accuracy of 89%. Alternatively, one can use the labels obtained by our unsupervised partitioning algorithm to train a supervised classifier. We evaluated two algorithms: An Artificial Neural Network (ANN) and a Random Forest (RF) classifier. Our ANN architecture consists of a collection of classifiers, each trained to detect one plankton class. The RF approach consists in a set of decision trees to separate the training step samples into the correct classes.

For comparison, a simple ANN classifier is trained using the labels provided by the unsupervised partitioning algorithm. The ANN is a massive parallel combination of single processing units which can learn the structure of the data and store the knowledge in its connections[24]. See Materials and Methods for further information and for a detailed description of the implemented architecture. The network is very shallow, providing an efficient feature selection process. The ANN classifier reaches a validation accuracy of 99% and a testing accuracy of 94.5%. Figure 3c,d report the ROC curves and the confusion matrix obtained by testing the trained ANN classifier on our ten species plankton dataset. The ROC curves are close to a perfect classifier and the confusion matrix is almost diagonal with minor overlap between two pairs of species: *Blepharisma americanuum-Paramecium bursaria* and *Spirostomum ambiguum–Stentor coerouleus*. This misclassification is primarily due to the similarity in the shape, size and texture of the two pairs of species, influencing both the unsupervised training clustering and the subsequent testing of the supervised classifier.

An alternative classifier method employs a Random Forest (RF) approach, a popular ensemble learning method used for classification and regression tasks.

We train an RF algorithm using the labels provided by the unsupervised classifier and reach an accuracy of 94%. For comparison, we train the same RF algorithm using the actual labels (ground truth) of the training set and reach an accuracy around 98%, proving that our unsupervised classification approach performs comparably well with respect to the correspondent supervised approaches for the trained classifier. Since the ANN performs marginally better than the RF classifier, we propose the former for a pipeline.

Anomaly detector. When deployed in the field, microscopes will encounter species that have never been seen before, so it is essential that such samples are detected and correctly identified as anomalies. For a given class, a sample is considered an anomaly if the sample features are significantly different from the feature average for the class. Algorithms for anomaly detection based on the separation of the features space have been successfully used to identify the intrusion in computer networks for security purposes[25]. Two anomaly detectors are implemented and compared; a state of the art one-class SVM and a customized neural network we call a Delta-Enhanced Class (DEC) detector that combines classification with anomaly detection. The one-class SVM algorithm uses a kernel to project the data onto a multidimensional space and can be interpreted as a two class SVM assigning the origin to one class and the rest of the data to another class. It then solves an optimization problem determining a hyperplane with maximum geometric margin, i.e., a surface where the separation between the two sets of points is maximal, that will be used as decision rule during the testing step.

A customized one-class SVM is implemented by normalizing the testing samples using the training data belonging to a single class. In this way, there will be a significant difference in the absolute value obtained for the anomaly (out-of-class) samples compared to the in-class samples, improving the accuracy of the SVM. The one-class SVM so designed reaches an average testing accuracy of (93.5 ± 6.0) %, with high accuracy in both anomaly detection and classification.

We now describe an alternative ANN-based approach that simultaneously performs classification and anomaly detection. As demonstrated above, a single layer ANN is able to satisfactorily classify plankton data from our in-house dataset. However, to effectively approach the anomaly detection step, we designed a deep neural network called Delta-Enhanced Class (DEC) detector (see materials and methods for further details). One DEC

Class	Number	Description
Geometric features	14	Area(pixels), Area (0-th order moment), perimeter, eccentricity, rectangularity, roundness, shape factor, width and height (minimum fitting rectangle), circularity, major and minor axis (fitting ellipse), equivalent diameter, convexity
Hu moments	7	Hu moments computed from normalized central image moments
Zernike moments	25	Zernike moments up to order 5
Image Intensity Features	8	Blue/green channels ratio, red/green channels ratio, red/blue channels ratio, gray levels histogram statistical features (skewness, kurtosis, mean value and standard deviation, entropy)
Haralick features	13	The first 13 Haralick descriptors computed from the Gray Scale Co-occurrence Matrix (GSCM)
Local binary patterns	54	Local binary patterns summarize structures of the image comparing each pixel to its neighborhood
Fourier descriptors	10	Fourier descriptors are contour-based features invariant with respect to rotation, translation and scaling

Table 1. List of morphological features extracted from the processed images. See Supporting Information for a detailed explanation.

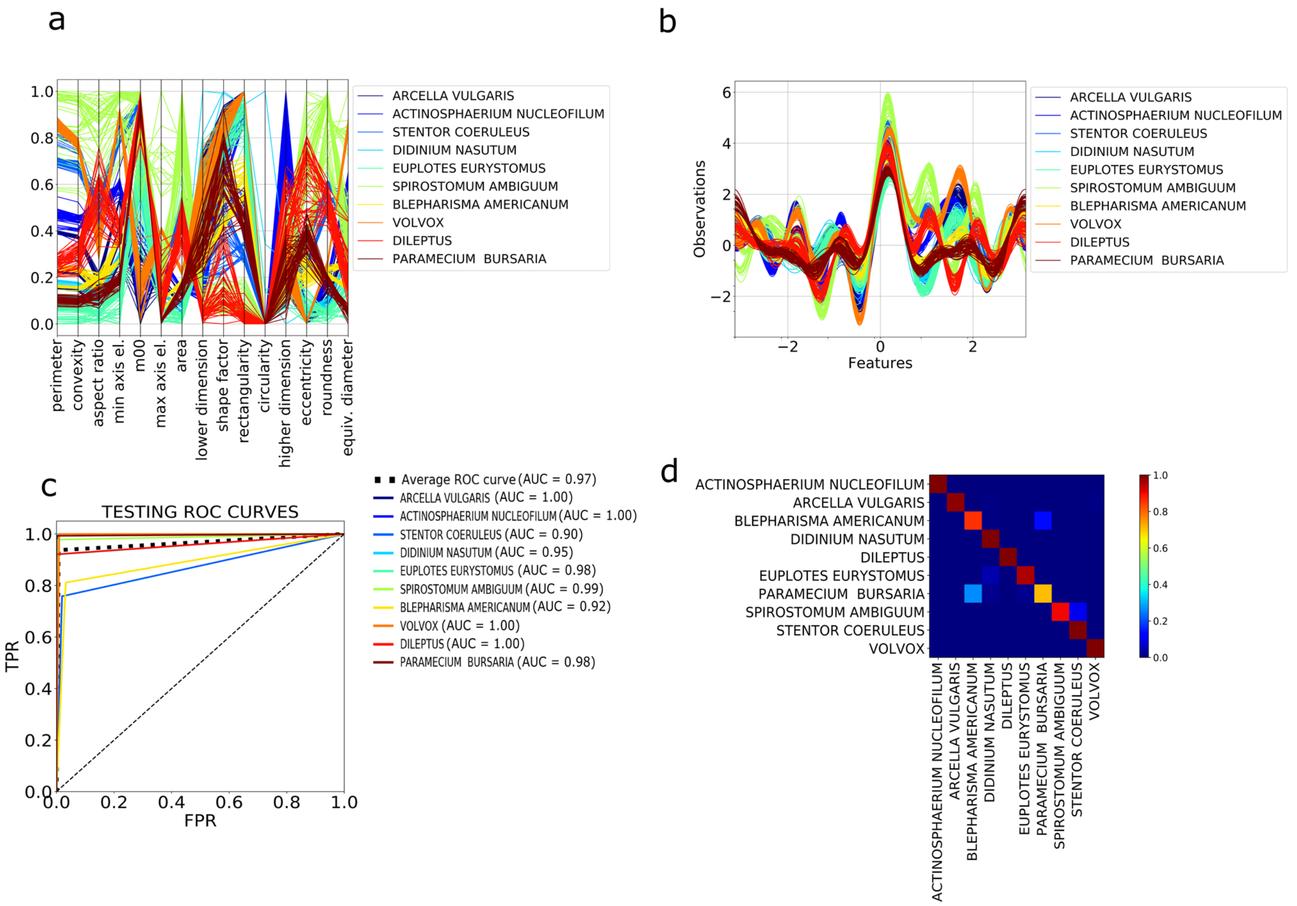
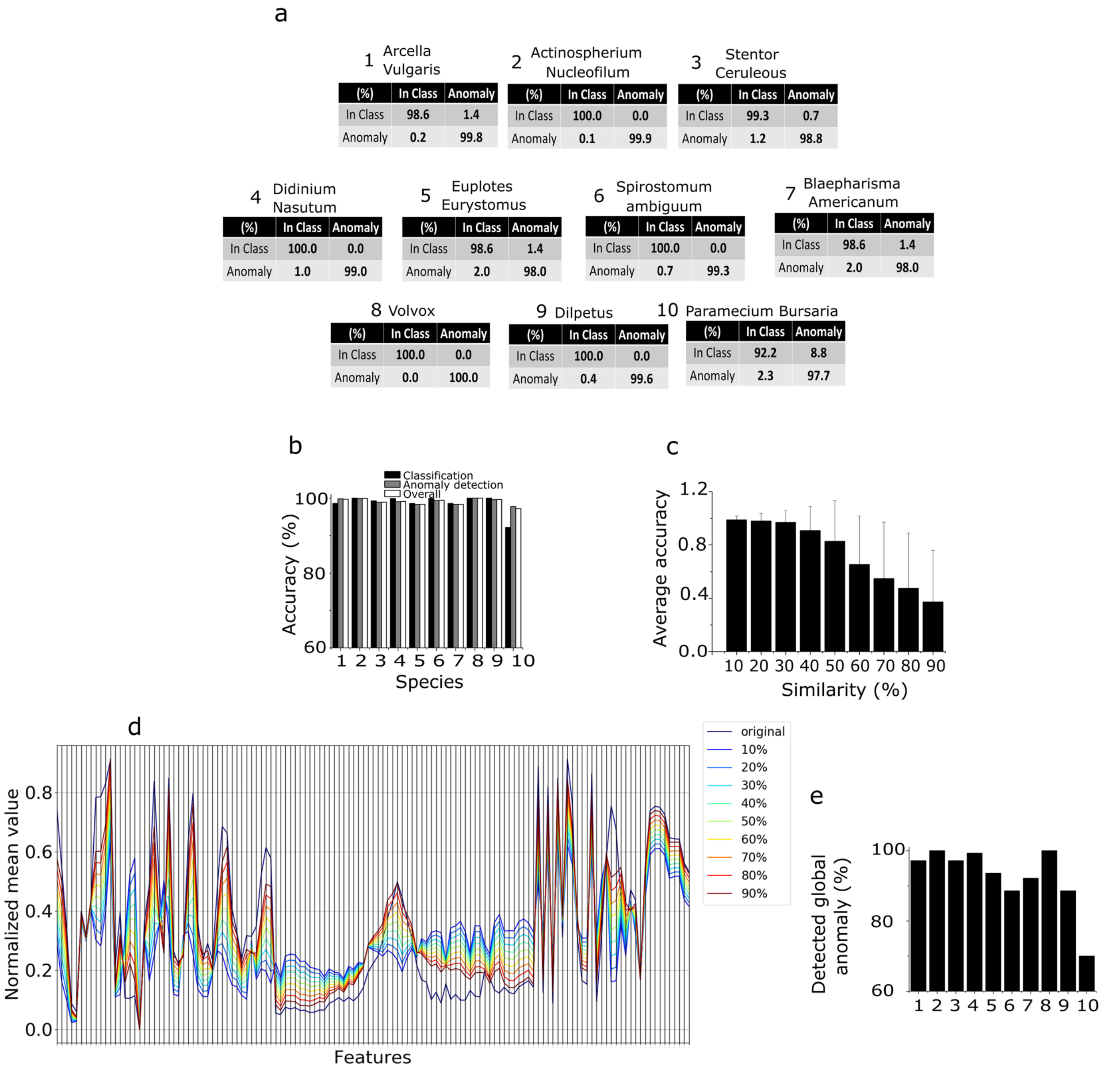


Figure 3. Feature space representation and classification performances. **a**, **b** Multidimensional visualization of the geometric subset of the ten species in the lensless microscope dataset, obtained using the following methods (see Supporting Information): **a** Andrew's curve. **b** Parallel coordinates. **c** ROC curves obtained for the neural network classifier trained on the labels provided by the clustering algorithm for the lensless microscope dataset. **d** Corresponding confusion matrix.

detector must be trained for each of the training species. Therefore, we train ten DEC detectors, one for each of the species of plankton identified in the unsupervised learning step. This procedure affords excellent accuracy on both classification and anomaly detection, on both real and simulated plankton data (see Fig. 4), with an average testing accuracy on real data of 98.8 ± 2.4%, an average anomaly detection testing accuracy of 99.2 ± 0.7% and an average overall testing accuracy of 99.1 ± 0.9% (see Fig. 4b for details). The confusion matrices in Fig. 4a demonstrate the discrimination power of our algorithm. The DEC detector outperforms the alternative one-class SVM classifier in both supervised (average accuracy equal to 95%) and unsupervised (average accuracy equal to 93.5%) configurations. It is worth reporting that the unsupervised one-class SVM reached a minimum overall accuracy of 79%, compared to 97.2% for the DEC detector (minimum values correspond to *Paramecium bursaria* detector). To test the overall performance of our method, we produce a dataset of surrogate plankton organisms.



1 Arcella Vulgaris

(%)	In Class	Anomaly
In Class	98.6	1.4
Anomaly	0.2	99.8

2 Actinosphenum Nucleofilum

(%)	In Class	Anomaly
In Class	100.0	0.0
Anomaly	0.1	99.9

3 Stentor Ceruleous

(%)	In Class	Anomaly
In Class	99.3	0.7
Anomaly	1.2	98.8

4 Didinium Nasutum

(%)	In Class	Anomaly
In Class	100.0	0.0
Anomaly	1.0	99.0

5 Euplotes Eurystomus

(%)	In Class	Anomaly
In Class	98.6	1.4
Anomaly	2.0	98.0

6 Spirostomum ambiguum

(%)	In Class	Anomaly
In Class	100.0	0.0
Anomaly	0.7	99.3

7 Blaepharisma Americanum

(%)	In Class	Anomaly
In Class	98.6	1.4
Anomaly	2.0	98.0

8 Volvox

(%)	In Class	Anomaly
In Class	100.0	0.0
Anomaly	0.0	100.0

9 Dilpetus

(%)	In Class	Anomaly
In Class	100.0	0.0
Anomaly	0.4	99.6

10 Paramecium Bursaria

(%)	In Class	Anomaly
In Class	92.2	8.8
Anomaly	2.3	97.7

Figure 4. Delta-enhanced class detector performances and results. **a** Confusion matrix corresponding to each of the ten neural networks trained on the lensless microscope dataset. **b** Overall testing accuracy performances for each of the ten testing classes. The number used on x axis to label each species correspond to the species number in panel **a**. **c**,**d** DEC detector anomaly detection performances tested on in silico generated data. **c** Testing accuracy performances for varying percentage values of in silico species similarity with the trained species. **d** Example of average features space parallel coordinates plot for the in-silico species obtained using the species *Spirostomum Ambiguum*. By increasing the similarity, the features of the surrogate species approach the features of the real species, resulting in an increased average anomaly misclassification rate, decreasing the overall accuracy levels. **e** Detection of unknown species. The panel shows the percentage of samples detected by all the DEC detectors as anomaly, when removing one training species from the set, for each of the ten training species. These numbers reflect the level of accuracy of the proposed algorithm in detecting unseen species. The number used on x axis to label each species correspond to the species number in panel **a**.

For each different species, we test the corresponding DEC detector architecture using a surrogate species created with a feature-by-feature weighted average of all the species in our dataset. Starting with a uniform weight distribution, we increase the weight for the species corresponding to the trained DEC detector architecture up to 0.9 (steps of 0.1), obtaining 9 different surrogate species (see Fig. 4d for an average parallel coordinates plot, showing the resulting distributions for the species *Spirostomum ambiguum*). The aim of this robustness test is to simulate the acquisition of an unknown species, whose features are increasingly closer to the features of the class correspondent to the detector, up to a maximum of 90% similarity. As Fig. 4e shows, our classifier can recognize the synthetic species as an anomaly with an average accuracy higher than 98% if the similarity between the synthetic and the real species is up to 30%, and it can maintain an average accuracy of over 82.6% if the species similarity is up to 50%. Accuracy of anomaly detection severely decreases if the species similarity is over 50%, reaching the minimum value of 37.5%.

Plankton classifier performance on the WHOI dataset. The WHOI provides a public dataset comprising millions of still monochromatic images of microscopic marine plankton, captured with an optical Imaging FlowCytobot (https://mclanelabs.com/imaging-flowcytobot/). To use this dataset as a benchmark to test our unsupervised classifier, we extract a set of 128 features from a collection of 40 species of plankton (100 images per species, randomly selected), using both the segmented binary image and the portion of the gray-scale image containing the plankton cell body. A full description of the species selection process is reported in the Supporting Information. The features set is identical to the one used for the lensless microscope dataset, except for the absence of three-color features, as the lensless microscope is a color-based sensor, while the Imaging FlowCytobot is monochromatic. Figure 2c, d show the results of our pipeline applied on the normalized features set. The algorithm reaches an overall purity value of 0.715 for the 40 WHOI species that we selected. The ability of our pipeline to distinguish between inter-species plankton morphology can be further observed comparing Fig. 2c, which represents the PCA space corresponding to a subset of 18 of the 40 species for the ground truth dataset, and Fig. 2d, which represents the corresponding PCA space resulting from the unsupervised partitioning algorithm. A complete PCA representation for the 40 species can be found in Supporting Information. We trained a random forest algorithm using the labels provided by the unsupervised partitioning with a train-test ratio of 80:20, obtaining a classification accuracy around 63%. We have also trained a supervised random forest algorithm using the ground truth labels on the extracted features, obtaining a classification accuracy around 79%. As a comparison, in[17] the authors obtained a classification accuracy around 87% on a fewer species subset extracted from the WHOI dataset (22 species), adopting a supervised SVM-based approach. In[20], the authors obtained a classification accuracy around 88% using features obtained with multiple learning kernels, on the same dataset.

The plankton classifier can reveal unseen species. We have demonstrated that our DEC neural networks are able to classify a sample as either a training class (i.e., the plankton species used to train the detector) or as an anomaly. If a sample is discarded by all the implemented detectors, it could either represent an intra-species anomaly (i.e., species included into the training set) or a sample belonging to an unseen species (i.e., species not included in the training set). The former represents the basis for using the proposed pipeline for real-time environmental monitoring, and its implications are discussed in the next section. We now test the potential of our pipeline to detect new species. We remove one class from our unsupervised partitioning ensemble set, consider it as never before seen and compute the number of testing samples detected as anomaly by all the remaining DEC detectors. This number indicates the algorithm accuracy in detecting new species. We repeat the procedure for each class. The average detection accuracy is 98.3 ± 10.1% (see Fig. 4e), demonstrating the ability of the pipeline to detect the presence of a new species. If two or more unseen species are detected, they will be stored as anomalies. As this group of anomalies grows, a human expert may determine offline the actual labels for these new species, thus allowing a DEC detector to be trained for each new species. Alternatively, the samples corresponding to unseen species may be clustered and classified by the unsupervised partitioning step of our pipeline, reducing the number of new species that must be examined by a human.

Discussion

The plankton classifier described in this paper provides the foundation for a robust, accurate and scalable mean to autonomously survey plankton in the field. We have identified interpretable and non-interpretable image features that work with our algorithms to perform an efficient clustering and classification on plankton data using minimal supervision and with a performance accuracy comparable to supervised learning algorithms[17]. Instead of labeling thousands of samples, an expert need only identifying one member of cluster to label all the samples of the cluster.

We introduced a neural network that performs classification by learning the shape of the feature space and uses this information to identify anomalies. The network uses a novel unbiased methodology of feature-to-feature comparison of a test sample to a random set of training samples. While most of the existing classification methods require various degrees of user input, our method is automated, without sacrificing performance accuracy or efficiency.

All features the plankton classifier relies upon are extracted from static images. However, our custom lensless microscope captures 2D and 3D dynamic of plankton. While this dynamic information is not considered in the analysis presented here, motion data can increase the dimensionality of the feature space, by adding spatio-temporal "behavioral" components, and may improve the performance of classifiers and anomaly detectors. This is particularly valuable in cases where species have considerable overlap in morphology feature space, as seen with *Blepharisma americanuum* and *Paramecium bursaria*, and *Spirostomum ambiguum* and *Stentor coerouleus*, shown in the confusion matrices in Fig. 3d. Currently, existing large plankton datasets, like the WHOI used in our validation experiments, are based on static images, but as the cost of video-based in situ microscopes drops and their deployment increases, we believe datasets that include spatio-temporal data will become available and the use of such features will gain importance.

Deploying smart microscopes capable of real-time continuous monitoring will give biologist an unprecedented view of plankton in situ. The adoption of an unsupervised unbiased pipeline is a significant step ahead in the development of a real-time "smart" detector for environmental monitoring. Several high-resolution acquisition systems for real-time plankton imaging already exist[26] and could adopt the pipeline proposed into this paper. Figure 5 shows a high-level representation of a continuous environmental monitoring system in the form of a flow chart, showing an example of how the detector could be coupled to the computational pipeline we

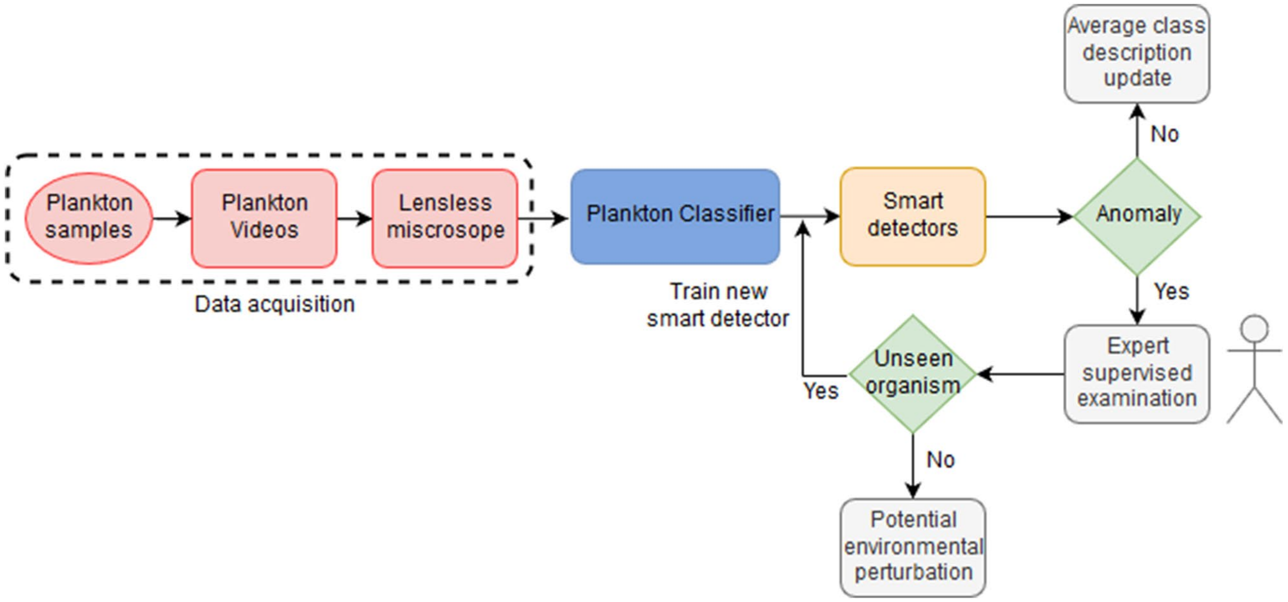


Figure 5. Proposed real-time smart environmental monitoring pipeline.

designed. Once the descriptors have been extracted from the acquired videos, it is possible to use them to build a set of DEC detectors. It is important to stress that the size of the data likely to be acquired, or already present in the databases, makes neural networks the obvious choice to carry out the analysis due to their unsurpassed scalability. Our newly designed and customized DEC detector neural architecture for plankton classification and anomaly detection is a functional and efficient example of such algorithm. Moreover, neural algorithms can infer non-linear relationships between features (input) and correlate them with the class description (output) without making any assumptions on the underlying learning model. Hence, the classification depends only on the extracted features. Every time the network identifies a species belonging to a specific class, the average set of morphological features is then updated, thereby further qualifying the class morphology phase space. If an anomaly is detected, it may be sent to an expert for a supervised examination. The expert will determine whether that sample could be a species not represented in the training set, or if it belongs to an existing training class, but its morphological features deviate significantly from the average features space of the corresponding class. In the former case, a new smart detector will be trained offline, so that the training set is dynamically expanded, and the system will provide a continuous monitoring of the aquatic environment using the human expert-in-the-loop paradigm. In the latter case, the identified anomalies may represent local environmental perturbations, either natural or man-made. Further work is needed to assess the validity of such hypothesis. An additional re-training step may be necessary to update the algorithms. Our pipeline is based on local analysis using a low powered device, capable of image capture and processing, classification and anomaly detection. Coupling such platform with a local (laptop, server) or cloud-based system where the training step may occur could provide the flexibility and resources needed to close the loop and generate the training data the low power platform can use for classification. Examples of systems that use this paradigm are already present in the literature[27], and we hope the availability of computational paradigms like the one we propose may increase the research in the field. Moreover, a desirable property of an unsupervised learning pipeline for classification is to be able to function across modalities of data acquisition. While our algorithms are optimized for the experimental apparatus we have developed, our results support wider applicability to datasets acquired with different instruments, with an accuracy not too far below more computationally taxing supervised machine learning methods. A high-resolution plankton acquisition system placed in the water and powered with our unsupervised pipeline may enable the development of real time continuous smart environmental monitoring systems that are fundamentally needed to stakeholders and decision-making bodies to monitor plankton microorganisms and, consequently, the entire aquatic ecosystem[28].

Finally, it is interesting to consider if such unsupervised approach can be utilized for different data types, thus widening the potential applicability and interest of the technique. While an extensive analysis of the performance of our pipeline on diverse set of data is beyond the scope of this work, it is worth commenting that the algorithms we use are general and pose no evident drawback to their application to other cell types. Particularly, the features our classifier uses to cluster the images do not include anything specific to plankton species (e.g. detection and estimation of number of flagella or other organelles.) Moreover, the proposed Deep Feature extraction method is even less dependent on the kind of data under study and may increase the applicability to other cell types. Thus, we expect the method to be potentially useful to other biological imaging fields.

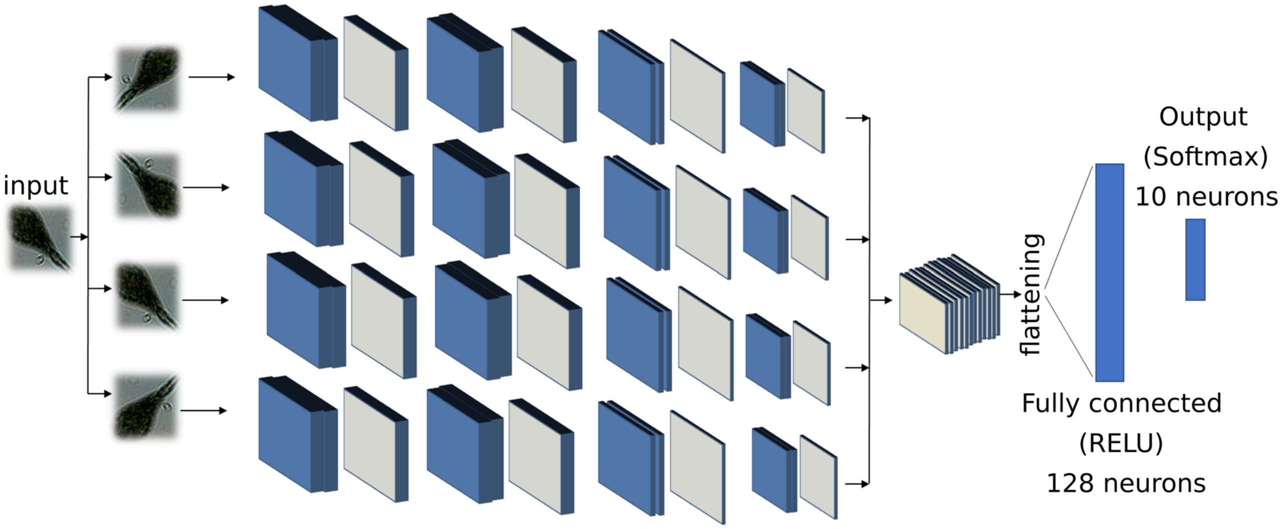


Figure 6. Deep features extraction. Deep CNN implemented for the purpose of deep features extraction. The blue layers represent convolutional layers, the grey ones represent a max pooling 2D operation. The fully connected layer with 128 neurons output has been used as feature set to the subsequent modules in our pipeline.

Material and methods

Description of the experimental apparatus and comparison with WHOI acquisition.

The lensless microscope[5] captures shadows of plankton swimming on top of an image sensor (OV5647, 3.76 × 2.74 mm, 2,592 × 1,944 pixels). Build with commodity components, the microscope does not require any optics or focusing adjustments, simplifying the construction, operation and cost of the device. To construct the microscope, the lens of a Raspberry Pi Camera was removed and replaced with PVC card (0.8 mm thick). A 12 mm square hold punched in the center of the PVC card provided a well for plankton samples. A single white LED was mounted at the top of a 100 mm black 38 mm diameter PVC tube, approximating a point light source, casting shadows of plankton onto the image sensor. Videos of plankton samples were recorded at HD 1,080p resolution (1,920 × 1,080 @ 30 fps).

The WHOI dataset was created using a submersible flow cytometer[29] The device produces shadow images (1,380 × 1,034) of individual plankton as they flow through a quartz imaging vessel. The flow cytometer produces higher resolution images than the lensless microscope. The flow cytometer illuminates plankton with a one-microsecond flash to minimize blurring and uses optical lenses to focus light rays onto an image sensor. The lenless microscope records video with continuous light source (non-strobed) with a minimum shutter speed of 200 microseconds. Diffraction is exacerbated by the distance between the plankton and the image sensor (~ 2 mm). The main advantage of the lensless microscope is cost, less than \$100 in materials, compared to ~ \$150,000 for a commercial flow cytometer[30].

Description of the pipeline.

The proposed unsupervised pipeline (i.e., the plankton classifier) shown in Fig. 1, consists of four modules: an image processor, a feature extractor, an unsupervised partitioning module and a classification module. In the following paragraphs we provide a description of the modules in more details, along with the methods considered and tested that led to our final design.

Image processing.

Each video consists of ten seconds of color video (1,920 × 1,080) captured at 30 frames per second. Background subtraction is applied to each frame to detect the swimming plankton in the image. A contour detector is applied to the processed image to create a bounding box around each plankter. Because of instrument design, organisms can swim in and out of the field of view (FOV) during acquisition. Our algorithm automatically selects organisms which are fully contained inside the FOV by checking whether the bounding box touches the borders of the FOV. In this way, the images we obtain will be only of fully visible organisms. The resulting cropped image is then saved. From this collection of images, a training set of 640 images (500 training and 140 testing) is selected for each class. An image processor module for static images has also been implemented for benchmarking the plankton classifier on existing plankton datasets (e.g., the WHOI dataset; See Supporting Information for further details.).

Feature extraction.

For each plankter image, 131 features are extracted from four categories: geometric (14), invariant moments (32), texture (67) and Fourier descriptors (10). Geometric features include area, eccentricity, rectangularity and other morphological descriptors, that have been used to distinguish plankton by shape and size[17]. The invariant Hu[31] (7) and Zernike moments[32] (25) are widely used in shape representation, recognition and reconstruction. Texture based features encode the structural diversity of plankton. Fourier Descriptors (FD) are widely used in shape analysis as they encode both local fine-grained features (high frequency FD) and global shapes (low frequency FD). A full list of the features we have selected is reported in Table 1. These features span a 131-dimensional space, capturing the biological diversity of the acquired plankton images. Figure 3a, b demonstrate as an example, the discriminating power of the geometrical features for the ten evaluated species.

Convolutional neural network (CNN) for deep features extraction. We implemented a deep CNN using eight convolutional layers and two fully connected layers, as described in Fig. 6. We customized our architecture to be invariant with respect to rotation, similar to what has been done in[19]. Each input sample is rotated four times at multiples of 90 degrees, and all the tensors resulting from the features extraction module are concatenated and used to train the fully connected layers. The neural network has been trained for 60 epochs, using stochastic gradient descent with learning rate equal to 10^{-5}, using data augmentation by means of translation, zooming, and rotation. It is worth noticing that the implemented rotational invariance module actually performs a data augmentation operation, and it is indeed useful when partial training data are available.

Unsupervised partitioning. *Partition entropy (PE).* The partition entropy (PE) coefficient is defined as:

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij} * \log(u_{ij}) \tag{1}$$

The coefficient is computed for every j in $[0, K]$ and takes values in range $[0, log(K)]$. The estimated number of clusters is assigned to the index j^* corresponding to the maximum PE value, PE(j^*). The lower the PE(j^*), the higher the uncertainty of the clustering. We repeat this procedure ten times and obtain a distribution of j^*. Finally, the estimation of the number of clusters Z is the mode of this distribution.

Clustering accuracy. Clustering accuracy is evaluated using purity:

$$purity = \frac{1}{N}\sum_{k}\max_{j}\left|w_k \cap c_j\right| \tag{2}$$

where the class k is associated to the cluster j with the highest number of occurrences. A purity value of one corresponds to clusters that perfectly overlap the ground truth. Purity decreases when samples belonging to the same class are split between different clusters, or when two or more clusters overlap with the same species. We have implemented a purity algorithm capable of checking for these occurrences and automatically adapt to the correct number of non-overlapping clusters (see Supporting Information).

Classification algorithms. *Random forest.* Random forests (RF) is a popular ensemble learning method[33] used for classification and regression tasks, introduced in 2001 by Breiman. Random forests model providing estimators of either the Bayes classifier or the regression function. Basically, RF work building several binary decision trees using a bootstrap subset of samples coming from the learning sample and choosing randomly at each node a subset of features or explanatory variables[34]. Random forests are often used for classification of large set of observations. Each observation is given as input at each of the decision tree, which will output a predicted class. The model outputs the class that is the mode of the class output by individual trees[35].

Let us consider a set of observations $x_1, x_2, \ldots, x_n$, with $x \in R^m$. The decision tree is designed as follows: we extract N times from the set of training observations (with replacement), for each of the total number of decision tree. We specify the number of features m^* to consider for the tree growing, with $m^* \ll m$. For each of the nodes in the tree, the algorithm randomly selects m^* features and calculates the best split for that node. The trees are only grown and not pruned (as in a normal tree classifier[36]. The split's aim is to reduce the classification error at each branch. In detail, the algorithm considers an entropy-based measure trying to reduce the amount of entropy at each branch, selecting, with such a procedure, the best split. A possible choice is the Gini index:

$$G_m = \sum_{i=1}^{K} p_{im}\left(1 - p_{im}\right) \tag{3}$$

where G_m is the Gini Index for branch at level m in the decision tree, and p_{im} is the proportion of observations assigned to class i. Minimizing G_m, means to decrease the heterogeneity at each branch, i.e., a best split will correspond to a lower number of class in the children nodes. The algorithms continue in growing trees until convergence on the entropy-based on the generalization error[35].

Neural networks. An artificial neural network (or multi-layer perceptron) is a massive parallel combination of single processing unit which can acquire knowledge from environment through a learning process and store the knowledge in its connections[24]. Classification is one of the most active research and application areas of neural networks. In this work we used an artificial neural network to build a classifier able to predict the species for each observation extracted using the shadow microscope. Figure 2 shows the developed architecture. The network is very shallow, with two hidden layers of 40 neurons and an output layer with as much neurons as the number of species to classify. As reported in the main text of this manuscript, we used a training dataset with 10 species, thus the output layer is made up of k neurons, where k is the number of clusters obtained using the unsupervised clustering. As Fig. 7 shows, the developed NN uses RELU activation function and dropout to reduce the overfitting. The network was trained using 200 epochs, Root mean square as an optimizer, a learning

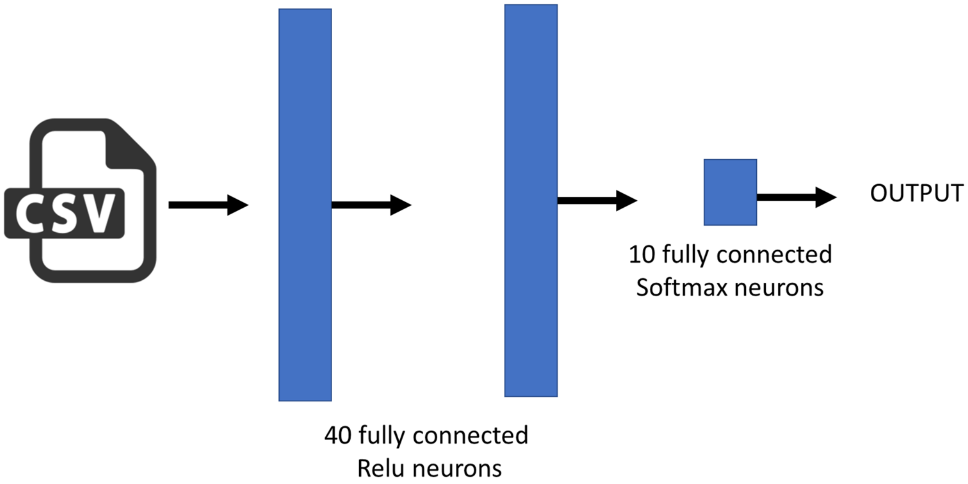


Figure 7. ANN architectures implemented for classification based on the extracted features.

rate $\lambda = 0.005$ and categorical cross-entropy as loss function. The training requires 50 s on a MAC book PRO, core i7—2.9 GHz, solid state disk and 16 GB of RAM. The neural network has been implemented using KERAS, a powerful high-level neural network API running on top of TensorFlow.

Anomaly detection. *One class SVM.* We adopted the one class SVM described by Scholpoff in[37]. Let us consider a set of N observations: $\left\{x_i, y_i\right| \in R^m, y_i = +1\}$. Where x_i is a m-dimensional real vector and $y_i = +1$ simply imply that the set contains normal observations belonging to a certain class. The one-class SVM is a classification algorithm returning a function which takes +1 in a "small" region capturing most of the data points, and −1 elsewhere. Let φ be a feature map that map our observations set x_i, into an inner product space such as the inner product for the image of φ can be evaluated using some simple kernel:

$$k(x, y) = \varphi(x)\varphi(y) \tag{4}$$

The strategy of the one class SVM is to map the data into the kernel space and separate the data from the origin with maximum margin, defining a hyperplane as:

$$H(x) = w * \varphi(x) - \rho \tag{5}$$

Meaning that we want to maximize the ratio $\frac{\rho}{\|w\|}$, corresponding to the hyperplane's distance from the origin. In order to solve this maximization problem, we have to solve a quadratic problem:

$$\begin{gathered} \min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu * m}\sum_i \xi_i - \rho \\ \text{subject to} \quad w * \varphi(x) \geq \rho - \xi_i, \xi_i \geq 0. \end{gathered} \tag{6}$$

where $\varphi(x)$ is the feature mapping function that maps observations x into a feature space, ξ_i is a slack variable for outlier that allows observations to fall on the other side of the hyperplane $\nu \in [0, 1)$ is a regularization parameter determining the bounding for the fractions of outliers and support vectors.

If w and ρ solve this problem, then the decision function:

$$f(x) = \operatorname{sgn}(H(x)) \tag{7}$$

will be positive for most of the training observation, while w will be still small. The parameter influences the trade-off between the reported properties. To solve the quadratic form, we can use Lagrangian multipliers, obtaining:

$$L(w, \xi, \rho, \alpha, \beta) = \frac{1}{2}\|w\|^2 + \frac{1}{\nu * m}\sum_i \xi_i - \rho - \sum_{i=1}^{m} \alpha_i(w * \varphi(x) - \rho + \xi_i) - \beta_i\xi_i \tag{8}$$

and set the derivatives with respect to w, ξ and ρ and expanding using the kernel expression yields:

$$f(x) = sgn\left(\sum_i a_i k(x_i, x) - \rho\right) \tag{9a}$$

$$\alpha_i = \frac{1}{\nu * m} - \beta_i \leftrightarrow 0 \leq \alpha_i \leq \frac{1}{\nu * m} \tag{9b}$$

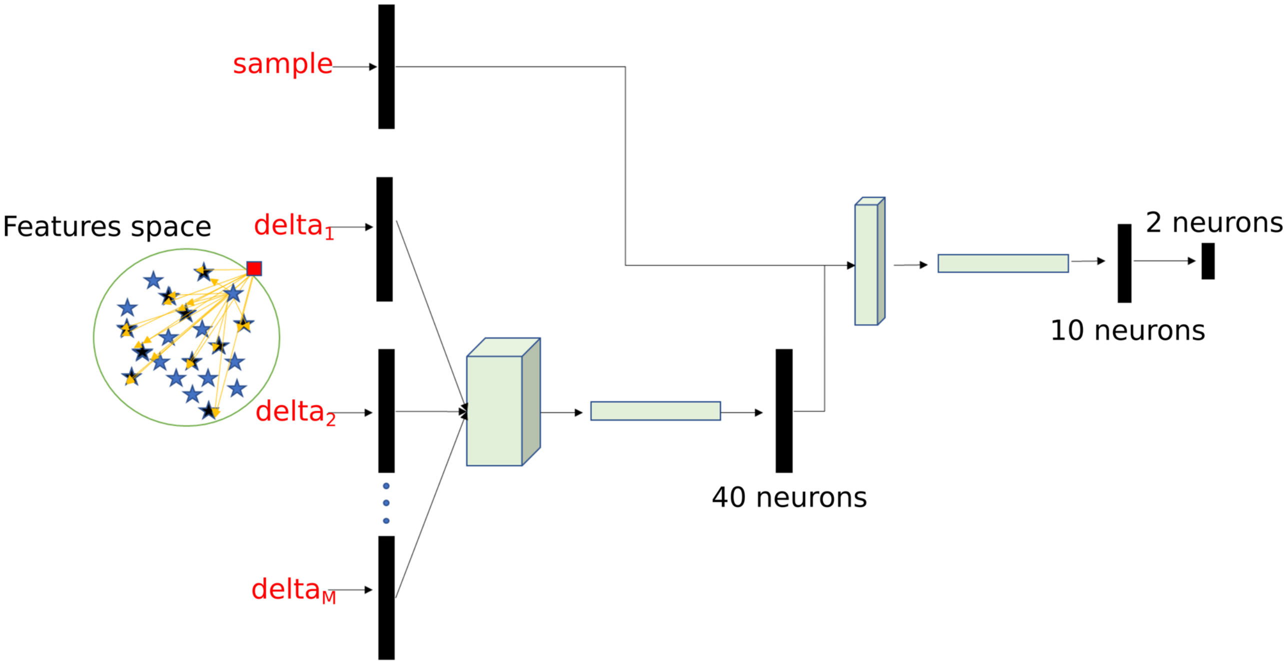


Figure 8. Schematic representation of DEC detector architecture.

$$\sum_{i=1}^{m} \alpha_i = 1 \tag{9c}$$

We used a Radial Basis Function kernel (RBF):

$$k(x_i, x) = e^{-\frac{\|x_i - x\|^2}{2\sigma^2}} \tag{10}$$

and then the original quadratic problem is solved substituting Eq. 9 into Eq. 8, yielding:

$$\min_{\alpha} \sum_{i=1}^{m} \alpha_i \alpha_j k\left(x_i, x_j\right) \tag{11}$$

under the constraint of Eqs. (9b) and (9c).

We finally use the support vectors x_i to recover the parameter ρ needed to compute the hyperplane:

$$\rho = w * \varphi(x) = \sum_{j} \alpha_j k\left(x_i, x_j\right) \tag{12}$$

DEC detectors. We designed a deep neural network that we named Delta-Enhanced Class (DEC) detector for the purpose of anomaly detection. The DEC detector's architecture is represented in Fig. 8, and shows a 2-neurons output, indicating that the sample is a member of the class or is an anomaly (i.e. not a member of the class). For each observation, we train such neural network with the actual features vector and extract randomly select a set of points from the training class in our dataset. For each of these selected points, we define a custom network layer (delta layer) that computes the difference in absolute value (as a vector, feature by feature) between the actual observation and the extracted random set. The vector of differences and the actual observations are used as inputs to the neural network (Fig. 8), which assigns the proper weights to either one during training. The set of points to select is a hyperparameter which needs to be tuned. Through testing we determine that 25 points is the optimal tradeoff accuracy and computational cost.

Code availability

The full source code accompanying this paper has been made available under EPL license at the following link: https://github.com/sbianco78/UnsupervisedPlanktonLearning.

Received: 13 January 2020; Accepted: 16 June 2020
Published online: 22 July 2020

References

1. Sournia, A., Chrdtiennot-Dinet, M.-J. & Ricard, M. Marine phytoplankton: How many species in the world ocean?. *J. Plankton Res.* **13**(5), 1093–1099. https://doi.org/10.1093/plankt/13.5.1093 (1991).
2. Behrenfeld, M. J. *et al.* Biospheric primary production during an ENSO transition. *Science* **291**(5513), 2594–2597. https://doi.org/10.1126/science.1055071 (2001).
3. Richardson, A. J. *et al.* Using continuous plankton recorder data. *Prog. Oceanogr.* **68**(1), 27–74. https://doi.org/10.1016/j.pocean.2005.09.011 (2006).
4. Fossum, T. O. *et al.* Toward adaptive robotic sampling of phytoplankton in the coastal ocean. *Sci. Robot.* **4**(27), eaav3041. https://doi.org/10.1126/scirobotics.aav3041 (2019).
5. Zimmerman, T. G. & Smith, B. A. Lensless stereo microscopic imaging. In *ACM SIGGRAPH 2007 Emerging Technologies, New York, NY, USA* (2007). https://doi.org/10.1145/1278280.1278296.
6. Sosik, H. M., Peacock, E. E., & Brownlee, E. F. *Annotated Plankton Images—Data Set for Developing and Evaluating Classification Methods*. https://doi.org/10.1575/1912/7341.
7. Schmid, M. S., Aubry, C., Grigor, J. & Fortier, L. The LOKI underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the Arctic Ocean. *Comput. Vis. Oceanogr.* **15–16**, 129–160. https://doi.org/10.1016/j.mio.2016.03.003 (2016).
8. Culverhouse, P. F., Ellis, R. E., Simpson, R. G., Williams, R., Pierce, R. W., & Turner, J. T. *Categorisation of five species of Cymatocylis (Tintinidae) by Artificial Neural Network*, Vol. 107, 273–280 (1994).
9. Orenstein, E. C. & Beijbom, O. Transfer learning and deep feature extraction for planktonic image data sets. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1082–1088 (2017) , https://doi.org/10.1109/WACV.2017.125.
10. Lumini, A. & Nanni, L. *Deep Learning and Transfer Learning Features for Plankton Classification* 51 (2019). https://doi.org/10.1016/j.ecoinf.2019.02.007.
11. Qiao, Hu. & Davis, C. Automatic plankton image recognition with co-occurrence matrices and support vector machine. *Mar. Ecol. Prog. Ser.* **295**, 21–31 (2005).
12. M. C. B. | D. of Oceanography *et al.*, *RAPID: Research on Automated Plankton Identification, Oceanography*, vol. 20 (2007). https://doi.org/10.5670/oceanog.2007.63.
13. Pastore, V. P., Zimmerman, T., Biswas, S. K. & Bianco, S. Establishing the baseline for using plankton as biosensor, *Presented at the Proceedings of the SPIE*, Vol. 10881 (2019). https://doi.org/10.1117/12.2511065.
14. Biswas, S. K. *et al.*, High throughput analysis of plankton morphology and dynamic, Presented at the Proceedings of the SPIE, Vol. 10881 (2019). https://doi.org/10.1117/12.2509168.
15. Schulze, K., Tillich, U. M., Dandekar, T. & Frohme, M. PlanktoVision—An automated analysis system for the identification of phytoplankton. *BMC Bioinform.* **14**, 115–115. https://doi.org/10.1186/1471-2105-14-115 (2013).
16. Dai, J., Wang, R., Zheng, H., Ji, G., & Qiao, X. *ZooplanktoNet: Deep Convolutional Network for Zooplankton Classification* 1–6 (2016). https://doi.org/10.1109/OCEANSAP.2016.7485680.
17. Sosik, H. M. & Olson, R. J. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnol. Oceanogr. Methods* **5**(6), 204–216. https://doi.org/10.4319/lom.2007.5.204 (2007).
18. Blaschko, M. B. *et al.*, Automatic in situ identification of plankton. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)—olume 1*, vol. 1, 79–86 2005. https://doi.org/10.1109/ACVMOT.2005.29.
19. Dieleman, S., De Fauw, J., & Kavukcuoglu, K. *Exploiting Cyclic Symmetry in Convolutional Neural Networks*, *ArXiv E-Prints*, arXiv:1602.02660 (2016).
20. Zheng, H. *et al.* Automatic plankton image classification combining multiple view features via multiple kernel learning. *BMC Bioinform.* **18**(16), 570. https://doi.org/10.1186/s12859-017-1954-8 (2017).
21. Hughes, A., Mornin, J. D., Biswas, S. K., Bauer, D. P., Bianco, S., & Gartner, Z. J. *Quantius: Generic, high-fidelity human annotation of scientific images at 105-clicks-per-hour*, bioRxiv, 164087 (2017). https://doi.org/10.1101/164087.
22. Reynolds, D. A. Gaussian mixture models. *Encycloped. Biom.* https://doi.org/10.1007/978-0-387-73003-5_196 (2009).
23. Romero, A., Gatta, C. & Camps-Valls, G. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **54**(3), 1349–1362. https://doi.org/10.1109/TGRS.2015.2478379 (2016).
24. Haykin, S. *Neural Networks: A Comprehensive Foundation* 1st edn. (Prentice Hall PTR, Upper Saddle River, 1994).
25. Bhuyan, M. H., Bhattacharyya, D. K. & Kalita, J. K. Network anomaly detection: Methods, systems and tools. *IEEE Commun. Surv. Tutor* **16**(1), 303–336. https://doi.org/10.1109/SURV.2013.052213.00046 (2014).
26. Zimmerman, T. *et al.*, Stereo in-line holographic digital microscope, Presented at the Proceedings of teh SPIE, Vol. 10883 (2019). https://doi.org/10.1117/12.2509033.
27. Grindstaff, B., Mabry, M. E., Blischak, P. D., Quinn, M. & J. C. Pires, *Affordable Remote Monitoring of Plant Growth and Facilities using Raspberry Pi Computers*, bioRxiv, 586776 (2019). doi: https://doi.org/10.1101/586776.
28. Scherer, C. *et al.*, *The Development of UK Pelagic Plankton Indicators and Targets for the MSFD* (2015).
29. Olson, R. J. & Sosik, H. M. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnol. Oceanogr. Methods* **5**(6), 195–203. https://doi.org/10.4319/lom.2007.5.195 (2007).
30. /ucscsciencenotes. https://ucscsciencenotes.com/feature/detecting-deadly-algae.
31. Huang, Z. & Leng, J. Analysis of Hu's moment invariants on image scaling and rotation. In *2010 2nd International Conference Computer Engineering Technology*, Vol. 7, V7–476-V7–480 (2010).
32. Yang, Z. & Fang, T. On the accuracy of image normalization by zernike moments. *Image Vis. Comput* **28**(3), 403–413. https://doi.org/10.1016/j.imavis.2009.06.010 (2010).
33. Ho, T. K. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition, 1995*, Vol. 1, pp. 278–282 (1995).
34. Genuer, R., Poggi, J.-M. & Tuleau, C. *Random Forests: some methodological insights*, ArXiv08113619 Stat (2008). Accessed Nov. 11, 2018. https://arxiv.org/abs/0811.3619.
35. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
36. Random forest algorithm for classification of multiwavelength data—IOPscience. https://iopscience.iop.org/article/10.1088/1674-4527/9/2/011. Accessed Nov. 11, 2018.
37. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural Comput.* **13**(7), 1443–1471. https://doi.org/10.1162/089976601750264965 (Jul.).

Acknowledgements

We thank Amanda K. Paulson and Aleksandar Godjoski for critical reading of the manuscript. We also thank all faculty and students in the National Science Foundation Center for Cellular Construction for discussion and critical feedback on the general idea and pipeline.

Author contributions

S.B. and T.Z. started the project, deciding to use the lensless microscope for acquiring real-time videos of swimming plankton using morphology and behavior to recognize any possible environmental threats. V.P.P acquired the lensless dataset, designed the pipeline, the A.I. system, wrote the code and implemented the methods for morphological analysis and testing, designed the DEC-detectors and wrote the paper, under the supervision of S.B. and T.Z. contributed in revising the paper and supporting the data acquisition process and the designing task. S.K.B. contributed in revising the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-68662-3.

Correspondence and requests for materials should be addressed to V.P.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.