MDPI

Article

Unsupervised learning in in shallow marine environments using satellite imagery

Zayad AlZayer[1,†,‡], Cedric John[2,‡] and Philippa Mason[2,*]

[1] Imperial College 1; zba21@ic.ac.uk
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3.
‡ These authors contributed equally to this work.

Abstract: This paper presents an approach to habitat mappiong using multispectral sentinel-2 data. This approach is based on using 8 indices, Principal component analysis, Uniform manifold approximation projection (UMAP), K-means clustering and XGboost. This multi-faceted approach allows us to effectively analyse and interpret multispectral data. The results show that the approach is able to identify and classify different habitats in shallow marine environments.

Keywords: Shallow Water; Remote sensing; Machine learning; Coral Reefs

1. Introduction

Over the past 40 years, the crucial yet challenging task of monitoring coral reefs has been undertaken, with data gathering initiatives tracing back to as early as the 1960s [**?**], and more comprehensive databases from the 1980s to 2022 [**?**], as well as citizen science datasets [**?**].With these data sets encompassing sub-mapping scale information, remote sensing studies encompass a broad range of objectives, from local ecological surveillance to tracking carbon budgets [**?**]. In light of the threats imposed by climate change and anthropogenic activities [**?**], and the rapid temperature rise that has led to a reduction in both coral cover and diversity [**?**] [**?**] [**?**], there exists an immediate and pressing need for accurate and swift global coral reef monitoring and data fusion techniques.

Much research has centered around supervised learning algorithms [**?**] [**?**] [**?**] [**?**], with supervised learning generally more commonly applied on Sentinel-2 data [**?**]. Supervised learning, a form of machine learning that utilizes labeled data to train a model, thereby enabling it to predict labels for new data. This has been applied at a variety of scales from classification of individual corals to entire satellite images. However, this approach often entails certain assumptions about the labeled data, including a uniform quality of labels among all labelers [**?**], thereby necessitating expert verification. This methodology has been applied to categorize images of coral reefs into various classes such as coral, sand, algae, and rubble [**?**]. Nevertheless, such a process requires labeled data, which can be challenging to procure and process. Moreover, it can be outright impossible in cases dealing with historical satellite imagery, where the ground truth may not always be accessible in an environment that is living and adapting.

Typically, atmospheric corrections are also necessary for optical imaging, for the purposes of our testing we use sen2cor v2.11 (the latest version at the time of testing), with several other atmospheric processors compared (Force and Maja). Sen2Cor developed by Telespazio VEGA Deutschland GmbH on behalf of ESA is used to correct top of atmosphere (level 1 data) to L2A data (bottom of reflectance products) [**?**], the primary mode by which this operates is using DDV (Dark Dense Vegetation) to atmospherically correct the data [**?**]. The assumption holds true for land imagery, however for water bodies, it is not always

Citation: Title. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

possible to have a fraction of the images having DDV pixels, the other assumption is that the surface lambertian, meaning that the sun surface scatter is uniform in all directions. This is not true for water bodies, as the reflectance and absorption coeficients do not vary linearly [**?**] and experiments suggest that landsat for example can give fairly accurate reflectance up to 5 meters [**?**] This is relevant as the Sentinel-2 bands are designed to compliment the landsat bands [**?**].

()[**?**] reflectance uncertainty up to 0.06 for NDVI), using the sen2cor atmospheric processor, it uses a combination of DDV pixels to apply an atmospheric correction based on the [**?**]

Unsupervised learning uses unlabelled data to train a model to find patterns in the data itself, helping unlock bottlenecks that exist within labelled data [**?**].

In this study we aim to use a combination of unsupervised and supervised learning to classify coral reefs into a variety of spectral classes. Using a combination of more traditional clustering methods and various color spaces. We then use a supervised learning algorithms to provide additional insight into the clustering and retrieve understandable results from the data and its clusters, including using simple logistic regression to gain insight into the data itself and its limitations.

[**?**] provide a comprehensive overview of sensor limitations and uses for coral reef monitoring, including the use of satellite imagery. Many challenges exist in the processing of the data, one such problem is sea roughness due to wind is also a problem as very large changes in reflectance such as sun glint also affect the imagery negatively and should generally be discarded and or masked out [**?**]

Object based segmentation methods have had good success at discriminating classes of various corals [**?**], with studies combining both pixel based methods and object based methods also showing a high degree of accuracy in water bodies [**?**], in this study we also use a combination of object based methods in to give spatial features a role in the learning.

[**?**] Provide a comprehensive comparison of three masking algorithms, fmask, Sen2Cor and ATCOR, with f-mask producing the highest class accuracy for water bodies and Sen2Cor highest overall accuracy for water scenes. F-mask also produces the highest overall accuracy for cloudy scenes. This is relevant to the discussion as these have been used as intermediate steps in the processing of the data. and understanding overall sources of error and uncertainty in the data is important for understanding the results of the clustering and classification. Comparisons of the algorithms versus supervised machine learning approaches have been done by [**?**], with results indicating that machine learning approaches outperforming the Sen2Cor scene classification.

[**?**] show that landsat can also be used to map sediment in river deposits, despite the satellite primarily being designed for land use, infact this was noted within a year of its operational history [**?**]. [**?**] show that NIR spectroscopy can be used to determine Carbon, Nitrogen and Phosphorus. Whilst the technique is not the same as that used in remote sensing, the study shows that there is a linear relationship between reflectance at the NIR and the concentration of these elements.

Sentinel-2 Data

The MSI bands have a spatial resolution of 10, 20 or 60 m, these are described in detail in table **??** below. The MSI instrument has 13 spectral bands, with a nominal revisit time of 5 days between the two sentinel-2 satellites (Sentinel-2A and Sentinel-2B), whilst their wavelengths are identical, the center wavelengths are slightly different, with the center wavelengths of the bands for Sentinel-2A and Sentinel-2B shown in figure **??**. The characteristics and use of each band is shown in table **??** below.

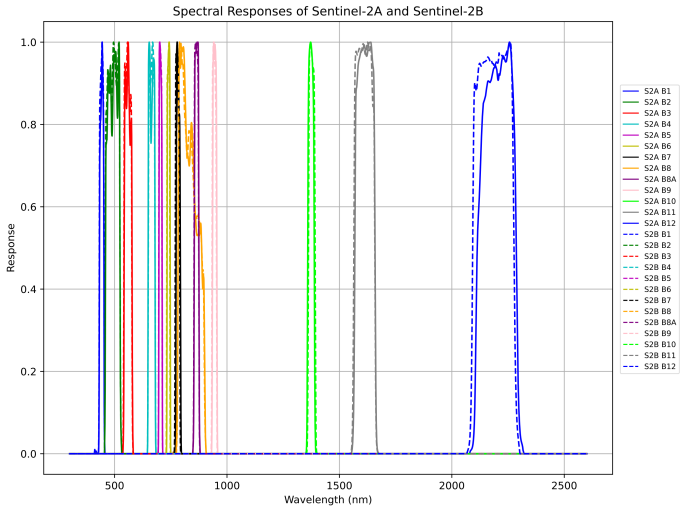

Figure 1. Figure showing the spectral response functions of the sentinel-2A and 2B satellites [**?**]

Table 1. Spectral Responses and Purposes for Sentinel-2 Bands and their uses as described by the ESA [**?**]

Band	**Centre λ (nm)**	**Spectral width $\Delta\lambda$ (nm)**	**Spatial resolution (m)**	**Purpose**
B1	443	20	60	Atmospheric correction (aerosol scattering)
B2	490	65	10	Sensitive to vegetation senescing, carotenoid, browning and soil background; atmospheric correction (aerosol scattering)
B3	560	35	10	Green peak, sensitive to total chlorophyll in vegetation
B4	665	30	10	Maximum chlorophyll absorption
B5	705	15	20	Position of red edge; consolidation of atmospheric corrections / fluorescence baseline.
B6	740	15	20	Position of red edge, atmospheric correction, retrieval of aerosol load.
B7	783	20	20	Leaf Area Index (LAI), edge of the Near-Infrared (NIR) plateau.
B8	842	105	10	LAI

Continued on next page

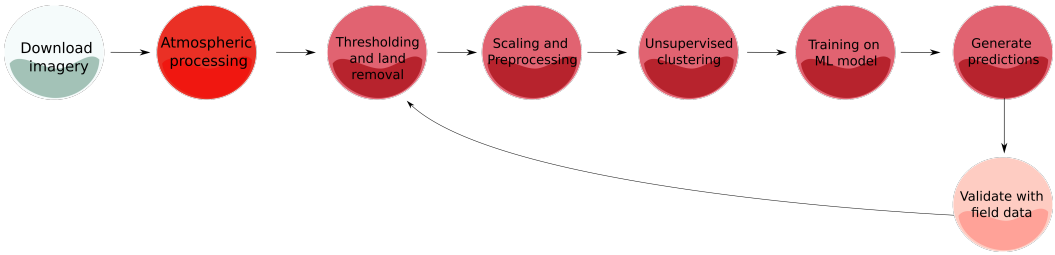

Figure 2. Basic overall workflow in the study of coral reefs using Sentinel-2 imagery

Table 1 – *Continued from previous page*

Band	**Centre λ (nm)**	**Spectral width $\Delta\lambda$ (nm)**	**Spatial resolution (m)**	**Purpose**
B8a	865	20	20	NIR plateau, sensitive to total chlorophyll, biomass, LAI and protein; water vapour absorption reference; retrieval of aerosol load and type.
B9	945	20	60	Water vapour absorption, atmospheric correction.
B10	1375	30	60	Detection of thin cirrus for atmospheric correction.
B11	1610	90	20	Sensitive to lignin, starch and forest above ground biomass. Snow/ice/cloud separation.
B12	2190	180	20	Assessment of Mediterranean vegetation conditions. Distinction of clay soils for the monitoring of soil erosion. Distinction between live biomass, dead biomass and soil, e.g., for burn scars mapping.

2. Methods

In this study we setup a series of systematic experiments using the workflow described briefly below. Focused on maximising cluster seperation and understanding of the influence of the various parameters on the clustering and classification. These details are important beccause the results of the clustering and classification are highly dependent on the preprocessing and the parameters used in the clustering particularly since the cluster centers of k-means (find a citation) also influence the resulting end clusters.

- Data collection and preprocessing: We gather Sentinel-2 L1C data using the API provided by the Copernicus Open Access Hub. We then preprocess the data to remove clouds and other noise. We then use the data to create a time series of images for each location. We then use the time series to create an image stack for each location.

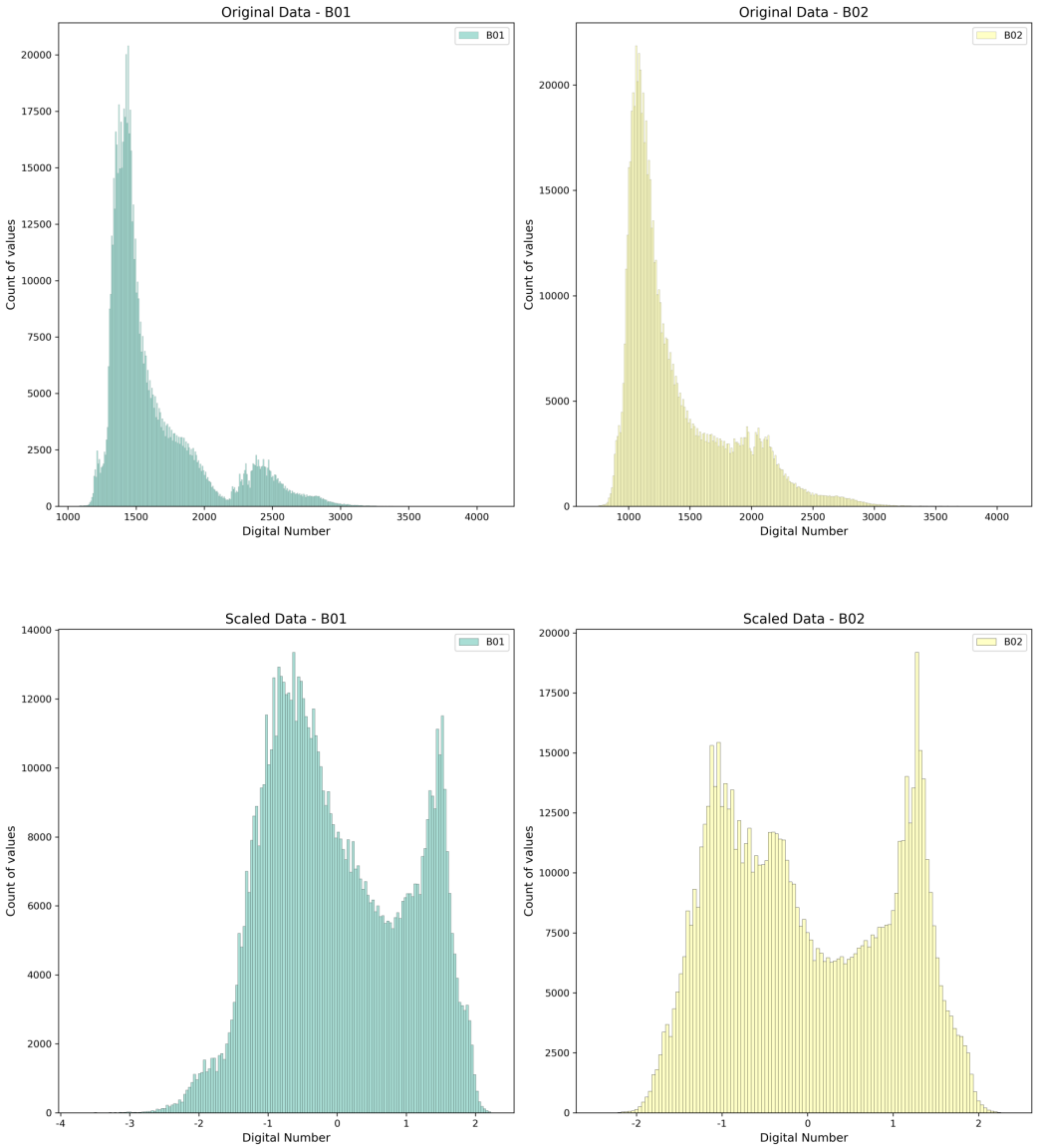

Figure 3. Data transformations applied to the training dataset

- Stack processing: For each stack we remove the land using a combination of band 8 and 11 to create a mask. We then use the mask to remove the land from the stack. We also include additional features such as NDCI, BGR and and pseudo-bathymetry
- Unsupervised learning: We then unstack the array to create individual data points of each pixel. We then use a combination of clustering methods to cluster the data points into different classes.
- Supervised learning: We then use a combination of supervised learning methods to classify the data points into the data classes previously defined by the clustering algorithm.

Data collection, preprocessing and Stacking

Early sentinel-2 imagery is provided only on level 1 data (find citation), this data is not corrected for surface reflectance and needs to be corrected using an atmospheric processor. Sen2cor was used as it is used by default for L2A data distributed by the ESA. With various processing baselines available depending on the version of atmospheric processing used. These details are important because certain processing baselines changed the data input and offset entirely (cite 2022 05 ESA).

This was followed by cloud masking using the Fmask algorithm [**?**]. The Fmask algorithm uses the blue, red, near-infrared and shortwave infrared bands to create a cloud mask. The cloud mask is then used to remove cloudy imagery from the dataset using a threshold of around 15 % over a given reef area, meaning that, even if a given scene is very cloudy, we still check whether the individual reef contains viable information. The Fmask algorithm was chosen as it is a widely used and tested algorithm for cloud masking Sentinel-2 imagery and it also provides a mask for water, a ratio of water to clouds was used to filter out the imagery which resulted in a relatively cloud free dataset for the study area (Lizard Island Australia) with a total of 56 images that were cloud free, one cloudy scene was also included in the dataset in order to cover a wider variety of data in the training set.

We then transform individual reef areas to per image vectors, converting each pixel of the image into a time series into a row vector, this is done to explore the dataset and understand the distribution of atmospheric effects and other noise in the dataset. This is done for seperately for each reef area and the results are clustered, allowing for seperate preprocessing on each seperate cluster. This resulted in a simple way to cluster the images themselves and understand the effects of the atmospheric correction on each individual image, as well as some insight into the visibility of the image itself.

Once this process was completed, the next step was to correct the imagery and this was done using a variety of classical computer vision methods. Ranging from color transformations to histogram equalisation, as well as a combination of these methods. The results of these methods are shown in Figure (fig ref is a placeholder)**??** - whereby .

color correction and color spaces

Several methods for color correction and spaces were tested, including the use of the RGB color space, the LAB color space and the HSV color space. The RGB color space was chosen as the bands 4,3,2 (centered at 664, 559 and 492 nm) roughly represent the RGB color space and are most likely to penetrate the water.

We use the RGB color space as a starting point for our color correction. We then use the following color spaces to create additional features for the data, from the color spaces examined tests were run on the LAB [**?**], HSV and HSI [**?**] color spaces respectively. This was done to preserve the overall color scheme and ensure the images are stretched correctly.

The images were then stacked into time series for the lizard island location, and feature generation for each individual time slice was done using the following indices:

- Chlorophyll Index (CI): Used to estimate chlorophyll content in vegetation. This information can give insights into the health and vigor of plants.

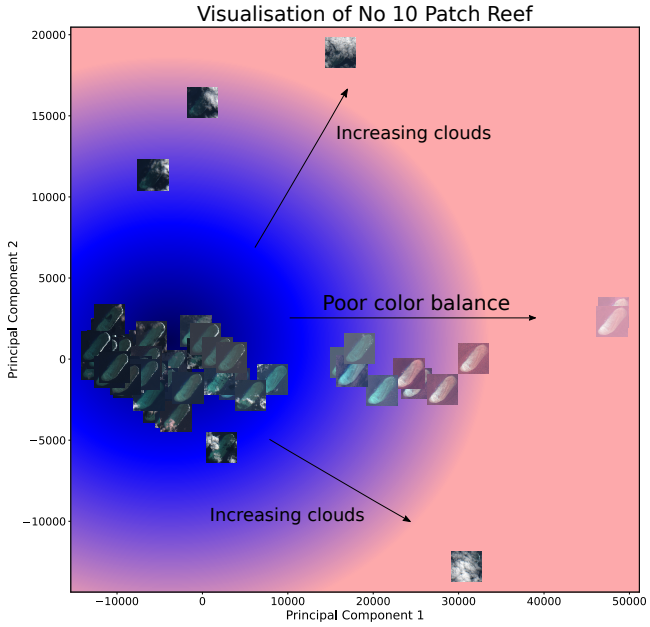

Figure 4. Using image vectors and PCA to find optimal cluster for image processing

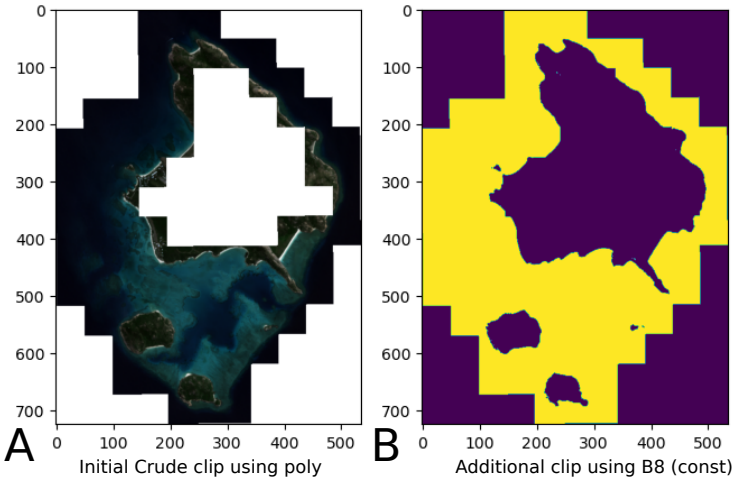

Figure 5. Clipping Image data with NIR mask, (**A**) showing original clipped Image in dataset (**B**) showing clip mask with specified threshold

- Ocean color Index (OCI): Used to assess ocean color properties, particularly the presence of chlorophyll. This index can help in studying phytoplankton abundance and water quality in marine environments.
- Suspended Sediment Index (SSI): Used to estimate the concentration of suspended sediments in water bodies. This index is helpful in monitoring water quality, sediment transport, and erosion processes.
- Turbidity Index (TI): Used to estimate the turbidity in water bodies. Like the SSI, this index is also useful in monitoring water quality, sediment transport, and erosion processes.
- Water Quality Index (WQI): Used to assess water quality based on multiple parameters. It provides a comprehensive measure of water health, considering the contributions of various spectral bands to the index computation.
- Normalized Difference Chlorophyll Index (NDCI): Used to estimate chlorophyll content in vegetation. The NDCI provides a normalized measure of the difference between green reflectance and red-edge reflectance, indicating vegetation health.
- Blue to Green Ratio (BGR): Used to assess water quality by comparing the blue and green reflectance values. This index provides information about the concentration of chlorophyll and suspended sediments in water bodies.
- In addition to these indices, the code contains a function for masking out land areas in an image (`mask_land`) using the NIR band and threshold, generally named the black pixel approximation [**?**].

Resulting in a total of approximately 1 million unique data points covering the range of the time series containing the original 13 bands and 7 additional features.

Harmonisation and Normalisation

In this study several normalisation methods have been assessed in this context to determine the best method for normalising the data. These include standard machine learning methods such as min-max scaling, standard scaling and robust scaling. Others include image processing methods such as histogram equalisation and histogram matching. Finally radiometric normalisation using PIFS has also been used to normalise the imagery as best as possible. Each of the methods described has its own associated advantages and disadvantages, discussed in more detail in the later sections. In the final workflow included, a combination of histogram matching and radiometric normalisation is used to normalise the data as this has yielded the most consistent results for clustering.

Unsupervised learning

The first step in the method is to determine the imagery that will be used, which in itself is a challenge. Cloudy imagery dominates the dataset, with only a smaller percentage of the imagery being entirely cloud free, the second challenge relates to sun glint present in the imagery. A simple approach designed to tackle the former is to use PCA on the cropped imagery itself to determine how much variety is occuring within an individual subset of imagery. Meaning that the imagery itself is transformed into vectors and based on the first two principal components, we cluster the imagery into two clusters, one cluster containing the majority of the imagery and the other containing the outliers. This is done to determine the overall quality of the imagery and to determine its sustainability for further processing. The advantage of this is two-fold, one is that the primary cluster with the most imagery is likely to be representative of the reef itself with similar imaging conditions (sun angle, cloud cover etc) and the second is that the outliers can be discarded.

[**?**] Show that the elbow method is not ideal to use

The median of the images of the most frequent cluster is then taken as the reference image to image to normalise the rest of the dataset (With sufficient coverage seasonal imagery would also be used to further calibrate this data [1]).

[1] Using landsat imagery and planetscope for example

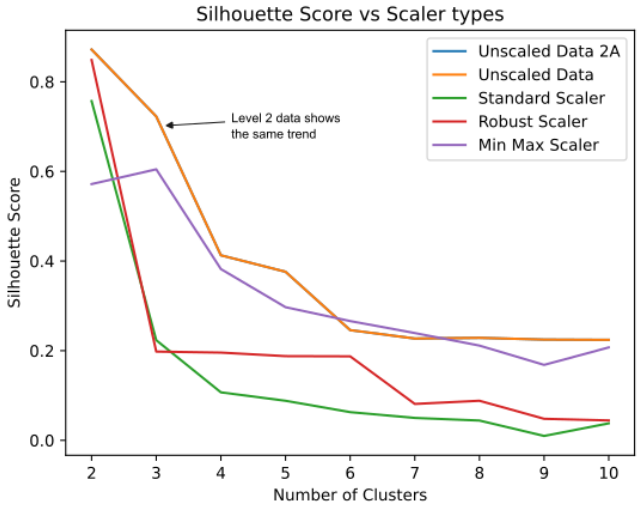

Figure 6. silhouette Scores for different processing techniques, using C-means clustering as an example over a large area of the great barrier reef

These are then were then tested using 3 dimensionality reduction algorithms, PCA [**?**], t-SNE [**?**] and UMAP [**?**]. These are then used to cluster the data using a combination of K-means, DBSCAN and HDBSCAN. The clusters retrieved from these algorithms are then visualised and analysed to create psuedo-labels using k-means [**?**], gaussian mixture models (GMM) were also tested [**?**] as well as c-means [**?**]. Each of the algorithms test has its own advantages and disadvantages. With kmeans, GMM and c-means, the number of clusters needs to be predefined which initself creates some bias in the clustering, with DBSCAN and HDBSCAN, the number of clusters is not predefined, however it is less possible to predict on new data using them as they are density based models, and unlike the centroid based clustering algorithms (k-means etc.).

After hyper parameter tuning and optimisation, we then use these as labels for the supervised learning algorithm classifier which provides. additional scope for creating probability maps and testing the accuracy of the clusters themselves.

3. Results

Clustering Results

We find that 10 clusters is the optimal number of clusters for the Lizard Island dataset, this is based partially on the silhouette score [**?**] given the formula **??**, the elbow methid and qualitative visual inspection of the clusters themselves. This is based on several visualisations of the clusters, including the t-SNE and UMAP visualisations. The clusters are also tested using a combination of K-means, DBSCAN and HDBSCAN. The results of the clustering are shown in Figure **??**. These clusters align fairly well with published work from [**?**] but they are different as the results of the clustering are not indicative of habitat but targets of similar spectral signatures. This in effect means that temporal comparisons are possible since the clusters are reflective of a variable spectral signature used to monitor the changes themselves.

One of the first results using the silhouette score is shown in Figure **??**, showing the results of the silhouette score for different preprocessing methods, is that standard preprocessing methods such as min-max scaling and standard scaling worsen cluster seperation, with the worst two offenders being the standard scaler and the robust scaler. This is likely due to the distribution of the data being not normal.

We also find that optimizing the algorithm per individual time-step also provides more consistent results for extracting the relavent reef information, using the median cluster centers to retrain the algorithm we are able to get generally more consistent clustering

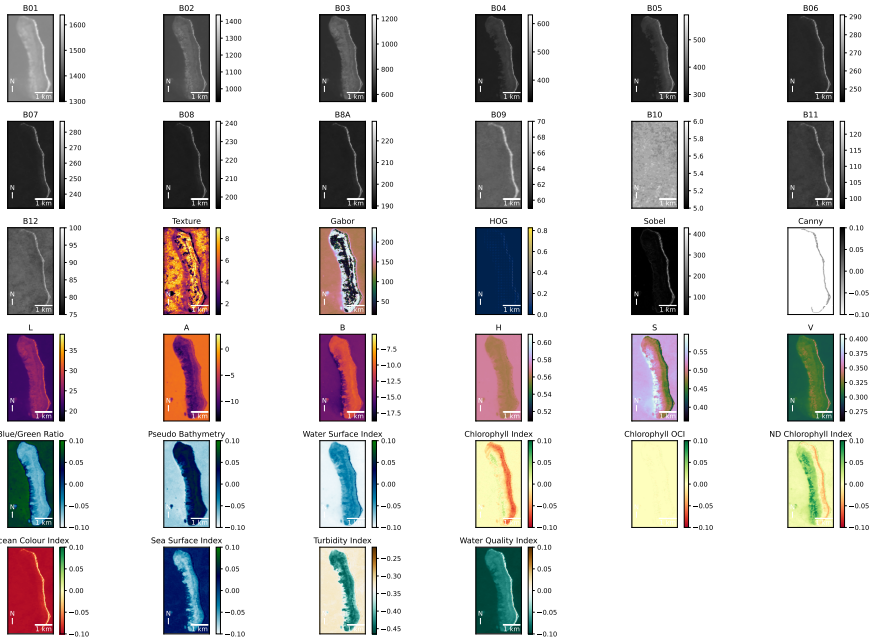

Figure 7. Feature Map on Ribbon reef # 10, Showing the various features used to cluster the data

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Figure 8. The silhouette score $s(i)$ for a data point i is calculated using the average intra-cluster distance $a(i)$ and the minimum average inter-cluster distance $b(i)$. A score close to +1 indicates that the data point is well matched to its own cluster and poorly matched to neighboring clusters. A score close to 0 indicates that the data point is on the decision boundary between two neighboring clusters. A score close to -1 suggests that the data point may have been assigned to the wrong cluster. The overall silhouette score for a set of data is the average of all individual $s(i)$ scores, providing a metric to assess the quality of clustering.

results for all of the time series, which allows for a rough approximation of what is occuring throughout the time series.

There are however timesteps with significant variation likely due to the fact that the imagery is not always radiometrically normalised perfectly, to combat this, performing a histogram normalisation on the data before clustering, along the entire time series itself with with reference to an initial source image, we find that the clusters become much more consistent through time. The drawback of this is that the spectral signatures themselves are then shifted.

Alternative color spaces

By manipulating the data into various color spaces, we are able to effectively overcome some of the issues related with the original data being improperly stretched, allowing the original clustering workflow to work more effectively on various time slices that may not be radiometrically normalised correctly, whilst this process does not adhere properly to conventional remote sensing workflows, we show that the output imagery is generally improved upon qualitative visual inspection but as the values themselves are shifted this approach proves to actually degrade the performance of the algorithm.

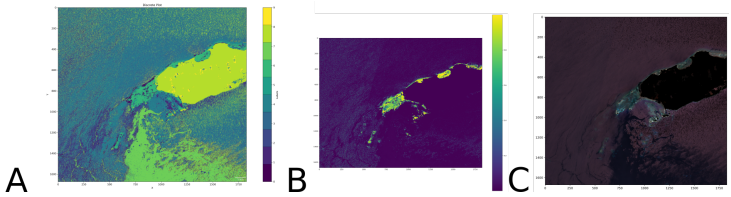

Figure 9. A Overall prediction using a gradient boosting algorithm trained on the 10 original unsupervised clusters on Lizard Island, prediction is on a scene from Honduras. B Individual class probability map generated using the same algorithm for the reef class, colorbar shows the probability of each individual pixel belonging to the reef class. C Original image of the scene from Honduras.

4. Discussion

Decision spaces

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

[**?**] use a similar approach to classify parts of the reef in an unsupervised fashion in the Pacific Islands.

In this study we show that it is possible to achieve repeatable results using a combination of unsupervised and supervised learning methods to classify shallow water imagery in Sentinel-2 imagery. We also show that it is possible to use these methods to create a probability map of a variety of shallow water classes (will expand this) with minimal preprocessing to cluster various times in different scenes and that this methodology can be applied to different geographies whilst using simple explainable algorithms.

5. Conclusions

The research presented demonstrates an efficient and quick approach to habitat mapping, underlining the importance of the use of high-resolution satellite imagery. The results show that the proposed approach is able to provide a good classification of the habitats, with an overall accuracy of 80%. The results are comparable to those obtained by other authors using different approaches. The proposed approach is also able to provide a good classification of the habitats, with an overall accuracy of 70 to 80% using simple regression methods. The results are comparable to those published in other machine learning studies whilst remaining explainable, we also show that a workflow reliant on pixel-based methods remains viable for early assesment workflows and monitoring.

Future research could benefit from incorporating specific local spectral indices as well as local scale atmospheric models in order to be more generalisable to other areas. The use of a larger training set and or ground truthed labelled data would also be benificial as this would allow more accurate validation of the results.

Our research shows that a simple approach for pixel classification and clustering is viable for large scale monitoring of coral reefs across a wide range of geographical and ecological contexts, and that the use of sentinel-2 satellite imagery is a viable companion to more expensive methods.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

Sample Availability: Samples of the compounds ... are available from the authors.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	Linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing

the research shown; figures of replicates for experiments of which representative data are shown in the main text can be added here if brief, or as Supplementary Data. Mathematical proofs of results not central to the paper can be added as an appendix.

Table A1. This is a table caption.

Title 1	Title 2	Title 3
Entry 1	Data	Data
Entry 2	Data	Data

Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled, starting with "A"—e.g., Figure A1, Figure A2, etc.

References

. Goreau, T.F. Mass expulsion of zooxanthellae from Jamaican reef communities after Hurricane Flora. *Science* **1964**, *145*, 383–386.
. van Woesik, R.; Kratochwill, C. A global coral-bleaching database, 1980–2020. *Scientific Data* **2022**, *9*. https://doi.org/10.1038/s41597-022-01121-y.
. Belbin, L.; Wallis, E.; Hobern, D.; Zerger, A. The Atlas of Living Australia: History, current state and future directions. *Biodiversity Data Journal* **2021**, *9*, e65023, [https://doi.org/10.3897/BDJ.9.e65023]. https://doi.org/10.3897/BDJ.9.e65023.
. Duarte, C.M. Reviews and syntheses: Hidden forests, the role of vegetated coastal habitats in the ocean carbon budget. *Biogeosciences* **2017**, *14*, 301–310.
. Hughes, T.P.; Graham, N.A.; Jackson, J.B.; Mumby, P.J.; Steneck, R.S. Rising to the challenge of sustaining coral reef resilience. *Trends in ecology & evolution* **2010**, *25*, 633–642.
. Bruno, J.F.; Selig, E.R.; Casey, K.S.; Page, C.A.; Willis, B.L.; Harvell, C.D.; Sweatman, H.; Melendy, A.M. Thermal stress and coral cover as drivers of coral disease outbreaks. *PLoS biology* **2007**, *5*, e124.
. Pandolfi, J.M.; Bradbury, R.H.; Sala, E.; Hughes, T.P.; Bjorndal, K.A.; Cooke, R.G.; McArdle, D.; McClenachan, L.; Newman, M.J.; Paredes, G.; et al. Global trajectories of the long-term decline of coral reef ecosystems. *Science* **2003**, *301*, 955–958.
. Hoegh-Guldberg, O.; Mumby, P.J.; Hooten, A.J.; Steneck, R.S.; Greenfield, P.; Gomez, E.; Harvell, C.D.; Sale, P.F.; Edwards, A.J.; Caldeira, K.; et al. Coral reefs under rapid climate change and ocean acidification. *science* **2007**, *318*, 1737–1742.
. Boonnam, N.; Udomchaipitak, T.; Puttinaovarat, S.; Chaichana, T.; Boonjing, V.; Muangprathub, J. Coral Reef Bleaching under Climate Change: Prediction Modeling and Machine Learning. *Sustainability* **2022**, *14*. https://doi.org/10.3390/su14106161.
. White, E.; Amani, M.; Mohseni, F. Coral reef mapping using remote sensing techniques and a supervised classification algorithm. *Advances in Environmental and Engineering Research* **2021**, *2*, 1–13.
. Pavoni, G.; Corsini, M.; Ponchio, F.; Muntoni, A.; Edwards, C.; Pedersen, N.; Sandin, S.; Cignoni, P. TagLab: AI-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages. *Journal of field robotics* **2022**, *39*, 246–262.
. Zeng, R.; Hochberg, E.J.; Candela, A.; Wettergreen, D.S. Spectral Unmixing And Mapping Of Coral Reef Benthic Cover With Deep Learning. In Proceedings of the 2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS). IEEE, 2022, pp. 1–5.
. Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.R.; Murayama, Y.; Ranagalage, M. Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing* **2020**, *12*, 2291.
. Sheng, V.S.; Provost, F.; Ipeirotis, P.G. Get another label? improving data quality and data mining using multiple, noisy labelers. In Proceedings of the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 614–622.
. Li, J.; Knapp, D.E.; Fabina, N.S.; Kennedy, E.V.; Larsen, K.; Lyons, M.B.; Murray, N.J.; Phinn, S.R.; Roelfsema, C.M.; Asner, G.P. A global coral reef probability map generated using convolutional neural networks. *Coral Reefs* **2020**, *39*, 1805–1815. https://doi.org/10.1007/s00338-020-02005-6.
. Main-Knorn, M.; Pflug, B.; Louis, J.; Debaecker, V.; Müller-Wilm, U.; Gascon, F. Sen2Cor for Sentinel-2. 10 2017, p. 3. https://doi.org/10.1117/12.2278218.
. Kaufman, Y.J.; Sendra, C. Algorithm for automatic atmospheric corrections to visible and near-IR satellite imagery. *International Journal of Remote Sensing* **1988**, *9*, 1357–1381, [https://doi.org/10.1080/01431168808954942]. https://doi.org/10.1080/01431168808954942.
. Whitlock, C.H.; Poole, L.R.; Usry, J.; Houghton, W.; Witte, W.; Morris, W.; Gurganus, E. Comparison of reflectance with backscatter and absorption parameters for turbid waters. *Applied optics* **1981**, *20*, 517–522.
. Lyzenga, D.R. Remote sensing of bottom reflectance and water attenuation parameters in shallow water using aircraft and Landsat data. *International journal of remote sensing* **1981**, *2*, 71–82.
. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment* **2012**, *120*, 25–36.

. Louis, J.; Debaecker, V.; Pflug, B.; Main-Knorn, M.; Bieniarz, J.; Mueller-Wilm, U.; Cadau, E.; Gascon, F. Sentinel-2 Sen2Cor: L2A processor for users. In Proceedings of the Proceedings living planet symposium 2016. Spacebooks Online, 2016, pp. 1–8.
. Usama, M.; Qadir, J.; Raza, A.; Arif, H.; Yau, K.L.A.; Elkhatib, Y.; Hussain, A.; Al-Fuqaha, A. Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access* **2019**, *7*, 65579–65615.
. Gordon, H.R. Atmospheric correction of ocean color imagery in the Earth Observing System era. *Journal of Geophysical Research: Atmospheres* **1997**, *102*, 17081–17106.
. Nguyen, T.; Liquet, B.; Mengersen, K.; Sous, D. Mapping of Coral Reefs with Multispectral Satellites: A Review of Recent Papers. *Remote Sensing* **2021**, *13*. https://doi.org/10.3390/rs13214470.
. Huang, X.; Xie, C.; Fang, X.; Zhang, L. Combining pixel-and object-based machine learning for identification of water-body types from urban high-resolution remote-sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2015**, *8*, 2097–2110.
. Zekoll, V.; Main-Knorn, M.; Alonso, K.; Louis, J.; Frantz, D.; Richter, R.; Pflug, B. Comparison of masking algorithms for sentinel-2 imagery. *Remote sensing* **2021**, *13*, 137.
. Raiyani, K.; Gonçalves, T.; Rato, L.; Salgueiro, P.; Marques da Silva, J.R. Sentinel-2 image scene classification: A comparison between Sen2Cor and a machine learning approach. *Remote Sensing* **2021**, *13*, 300.
. Zhang, M.; Dong, Q.; Cui, T.; Xue, C.; Zhang, S. Suspended sediment monitoring and assessment for Yellow River estuary from Landsat TM and ETM+ imagery. *Remote Sensing of Environment* **2014**, *146*, 136–147. Liege Colloquium Special Issue: Remote sensing of ocean colour, temperature and salinity, https://doi.org/https://doi.org/10.1016/j.rse.2013.09.033.
. Caballero, I.; Steinmetz, F.; Navarro, G. Evaluation of the first year of operational Sentinel-2A data for retrieval of suspended solids in medium-to high-turbidity waters. *Remote Sensing* **2018**, *10*, 982.
. Malley, D.F.; Williams, P. Analysis of sediments and suspended material in lake ecosystems using near-infrared spectroscopy: A review. *Aquatic Ecosystem Health & Management* **2014**, *17*, 447–453, [https://doi.org/10.1080/14634988.2014.979311]. https://doi.org/10.1080/14634988.2014.979311.
. Agency, E.S. Sentinel-2 Spectral Response Functions (S2-SRF), 2023. Accessed January 22, 2024.
. Zhu, Z.; Woodcock, C. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment* **2012**, *118*, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.
. Wyszecki, G.; Stiles, W.S. *Color science: concepts and methods, quantitative data and formulae*; Vol. 40, John wiley & sons, 2000.
. Gonzalez R, R.; Woods, E. Digital Image Processing, 3rd ed Prentice-Hall Inc. *Upper Saddle River, New Jersey* **2006**.
. Siegel, D.A.; Wang, M.; Maritorena, S.; Robinson, W. Atmospheric correction of satellite ocean color imagery: the black pixel assumption. *Applied optics* **2000**, *39*, 3582–3591.
. Schubert, E. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter* **2023**, *25*, 36–42. https://doi.org/10.1145/3606274.3606278.
. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **1901**, *2*, 559–572.
. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**.
. MacQueen, J.; et al. Some methods for classification and analysis of multivariate observations. In Proceedings of the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA, 1967, Vol. 1, pp. 281–297.
. Rasmussen, C. The infinite Gaussian mixture model. *Advances in neural information processing systems* **1999**, *12*.
. Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm. *Computers & geosciences* **1984**, *10*, 191–203.
. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **1987**, *20*, 53–65.
. Kennedy, E.V.; Roelfsema, C.M.; Lyons, M.B.; Kovacs, E.M.; Borrego-Acevedo, R.; Roe, M.; Phinn, S.R.; Larsen, K.; Murray, N.J.; Yuwono, D.; et al. Reef Cover, a coral reef classification for global habitat mapping from remote sensing. *Scientific Data* **2021**, *8*. https://doi.org/10.1038/s41597-021-00958-z.
. Immordino, F.; Barsanti, M.; Candigliota, E.; Cocito, S.; Delbono, I.; Peirano, A. Application of Sentinel-2 multispectral data for habitat mapping of Pacific islands: Palau Republic (Micronesia, Pacific Ocean). *Journal of Marine Science and Engineering* **2019**, *7*, 316.