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 [1], and more comprehensive databases from the 1980s to 2022 [2], as well as citizen science datasets [3].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 [4]. In light of the threats imposed by climate change and anthropogenic activities [5], and the rapid temperature rise that has led to a reduction in both coral cover and diversity [6] [7] [8], 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 [9] [10] [11] [12], 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 [13], 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 [14]. 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.

Unsupervised learning is a type of machine learning that uses unlabelled data to train a model to find patterns in the data itself, helping unlock bottlenecks that exist within labelled data [15].

In this study we aim to use a combination of unsupervised and supervised learning to classify coral reefs into different 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.

Citation: AlZayer, Z.; Mason, P.; John, C. Title. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

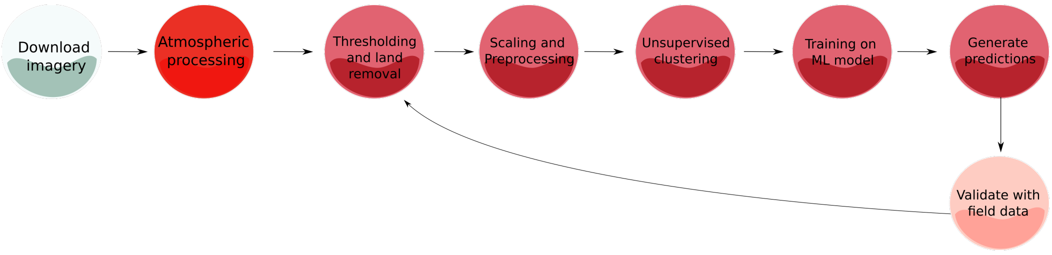

Figure 1. Basic overall workflow in the study of coral reefs using Sentinel-2 imagery

[**?**] 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 [16]

Object based segmentation methods have had good success at discriminating classes of various corals [17], with studies combining both pixel based methods and object based methods also showing a high degree of accuracy in water bodies [18]

2. Methods

In this study we setup a series of systematic experiments using the workflow described briefly below.

- 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.
- 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

Image correction for L1C data was done using the sen2cor processor for all the imagery to remove atmospheric effects. This was followed by cloud masking using the Fmask algorithm [19]. 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 the clouds from the image. 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 image was also included in the dataset in order to cover a wider variety of data in the training set.

Color correction and color spaces

As bands 4,3,2 (centered at 664, 559 and 492 nm) roughly represent the RGB color space, we use this 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 [20], HSV and HSI [21] color spaces respectively. This was done to preserve the overall color scheme and ensure the images are stretched correctly.

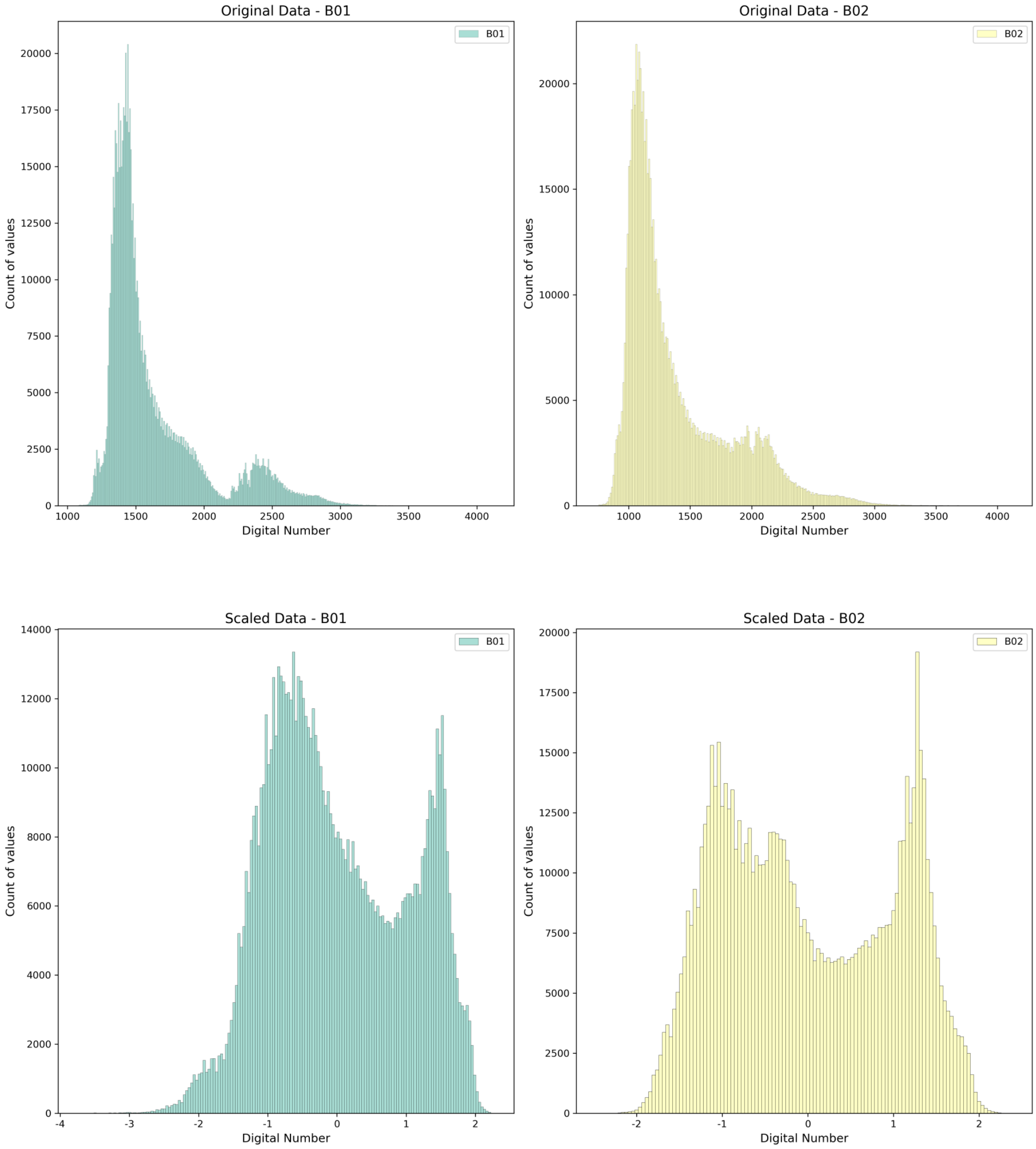

Figure 2. Data transformations applied to the training dataset

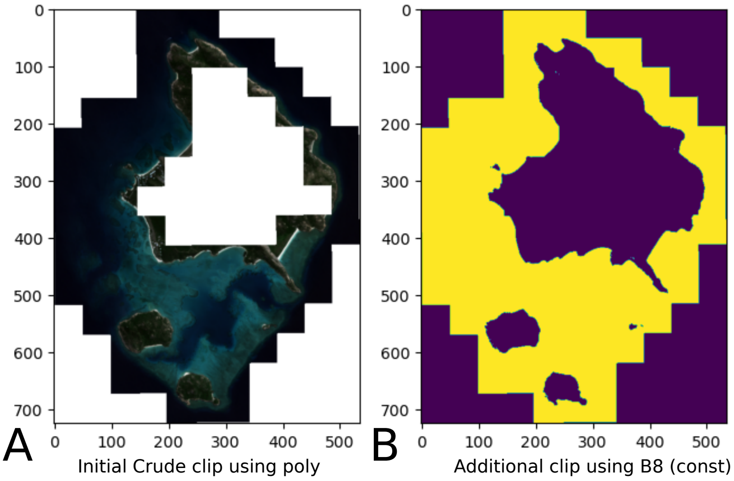

Figure 3. Clipping Image data with NIR mask, (**A**) showing original clipped Image in dataset (**B**) showing clip mask with specified threshold

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.
- 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 [22].

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.

Unsupervised learning

These are then entered into 3 dimensionality reduction algorithms, PCA [23], t-SNE [24] and UMAP [25]. These dimensionality reduction algorithms are then used to reduce the dimensionality of the data to 2 dimensions. 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 [26] and gaussian mixture models were also tested [27]. 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 on the silhouette score and the visual inspection of the clusters. 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 [28].

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

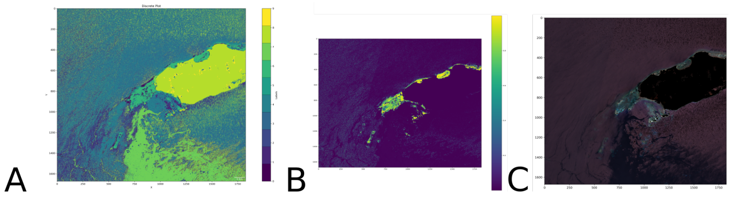

Figure 4. A Overall prediction using a gradient boosting algorithm trained on the 10 original unsupervised clusters on Lizard Island, prediction is on a scene from Honduras. B Individual class probability map generated using the same algorithm for the reef class, colorbar shows the probability of each individual pixel belonging to the reef class. C Original image of the scene from Honduras.

conventional remote sensing workflows, we show that the output imagery is generally improved upon qualitative visual inspection.

4. Discussion

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. 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.

Future research could benefit from incorporating specific local spectral indices 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

1. Goreau, T.F. Mass expulsion of zooxanthellae from Jamaican reef communities after Hurricane Flora. *Science* **1964**, *145*, 383–386.
2. 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.
3. 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.
4. Duarte, C.M. Reviews and syntheses: Hidden forests, the role of vegetated coastal habitats in the ocean carbon budget. *Biogeosciences* **2017**, *14*, 301–310.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. Gordon, H.R. Atmospheric correction of ocean color imagery in the Earth Observing System era. *Journal of Geophysical Research: Atmospheres* **1997**, *102*, 17081–17106.
17. 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.
18. 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.
19. 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.
20. Wyszecki, G.; Stiles, W.S. *Color science: concepts and methods, quantitative data and formulae*; Vol. 40, John wiley & sons, 2000.
21. Gonzalez R, R.; Woods, E. Digital Image Processing, 3rd ed Prentice-Hall Inc. *Upper Saddle River, New Jersey* **2006**.

22. 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.
23. 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.
24. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
25. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**.
26. 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.
27. Rasmussen, C. The infinite Gaussian mixture model. *Advances in neural information processing systems* **1999**, *12*.
28. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.