# Introduction

Coral monitoring has been a crucial and challenging task for over the past 40 years, with data collected as early as (1900 sometime - add citations). With studies collecting information about for a large variety of reasons ranging from local ecological monitoring to tracking carbon budgets (cite). With threats like climage change and human activity ((Hughes et al. 2010)) causing a decline in coral cover and diversity (cite), there is a need for rapid and accurate monitoring of coral reefs across the planet.

A great deal of research has been focused on supervised learning algorithms, a type of machine learning that uses labelled data to train a model to predict the label of new data. This has been used to classify coral reef images into different classes such as coral, sand, algae and rubble (cite). However, this requires a large amount of labelled data, which is often difficult to obtain and in some cases impossible when dealing with satellite imagery collected in the past where the ground truth is not always available.

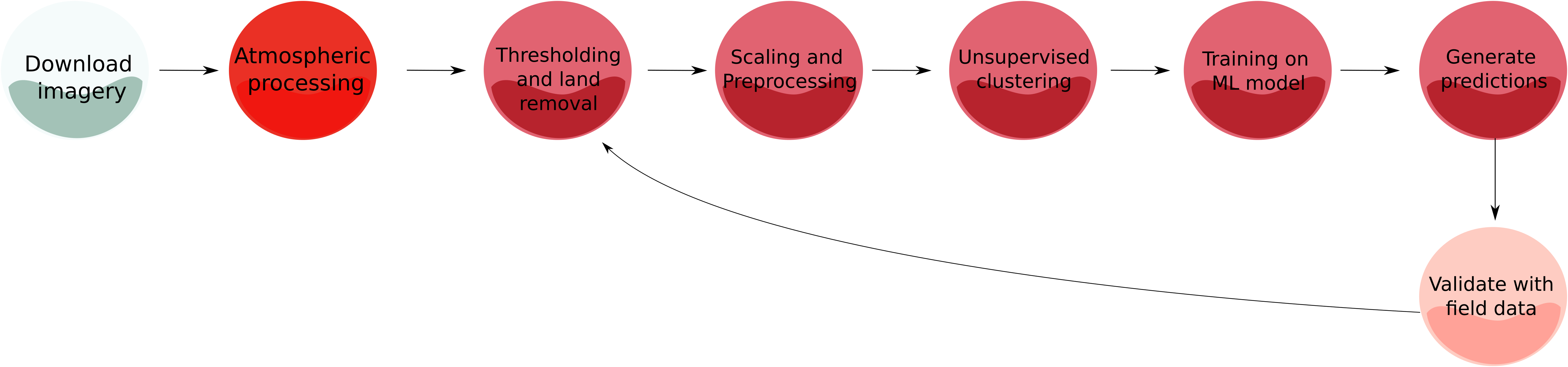
Unsupervised learning is a type of machine learning that uses unlabelled data to train a model to find patterns in the data itself.

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.

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 (Gordon 1997)

# Methods

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



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

* 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 (Zhu and Woodcock 2012). 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.

![](data:application/pdf;base64,)

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

## 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 (Wyszecki and Stiles 2000), HSV and HSI (Gonzalez R and Woods 2006) 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.
* 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 (Siegel et al. 2000).

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 (Pearson 1901), t-SNE (Van der Maaten and Hinton 2008) and UMAP (McInnes, Healy, and Melville 2018). 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 (MacQueen et al. 1967) and gaussian mixture models were also tested (Rasmussen 1999). 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.

# Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

# Discussion

## Subsection

### Subsubsection

Bulleted lists look like this:

* First bullet;
* Second bullet;
* Third bullet.

Numbered lists can be added as follows:

1. First item;
2. Second item;
3. Third item.

The text continues here.

## Figures, Tables and Schemes

All figures and tables should be cited in the main text as Figure [3](#fig1), Table [[tab1]](#tab1), etc.

|  |
| --- |
|  |

This is a figure. Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

CCC **Title 1** & **Title 2** & **Title 3**  
Entry 1 & Data & Data  
Entry 2 & Data & Data 1

The text continues here (Figure [4](#fig2) and Table [[tab2]](#tab2)).

|  |
| --- |
| -0cm image |

This is a wide figure.

-0cm

CCCC **Title 1** & **Title 2** & **Title 3** & **Title 4**  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data  
& Data & Data & Data

Text.

Text.

## Formatting of Mathematical Components

This is the example 1 of equation:

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

This is the example 2 of equation:

-0cm

Please punctuate equations as regular text. Theorem-type environments (including propositions, lemmas, corollaries etc.) can be formatted as follows:

Example text of a theorem.

The text continues here. Proofs must be formatted as follows:

*Proof of Theorem 1.* Text of the proof. Note that the phrase “of Theorem 1” is optional if it is clear which theorem is being referred to. ◻

The text continues here.

# 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.

# Conclusions

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

# Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

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.

CCC **Title 1** & **Title 2** & **Title 3**  
Entry 1 & Data & Data  
Entry 2 & Data & Data

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.

-0cm

Gonzalez R, RC, and E Woods. 2006. “Digital Image Processing, 3rd Ed Prentice-Hall Inc.” *Upper Saddle River, New Jersey*.

Gordon, Howard R. 1997. “Atmospheric Correction of Ocean Color Imagery in the Earth Observing System Era.” *Journal of Geophysical Research: Atmospheres* 102 (D14): 17081–106.

Hughes, Terry P, Nicholas AJ Graham, Jeremy BC Jackson, Peter J Mumby, and Robert S Steneck. 2010. “Rising to the Challenge of Sustaining Coral Reef Resilience.” *Trends in Ecology & Evolution* 25 (11): 633–42.

MacQueen, James et al. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–97. 14. Oakland, CA, USA.

McInnes, Leland, John Healy, and James Melville. 2018. “Umap: Uniform Manifold Approximation and Projection for Dimension Reduction.” *arXiv Preprint arXiv:1802.03426*.

Pearson, Karl. 1901. “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* 2 (11): 559–72.

Rasmussen, Carl. 1999. “The Infinite Gaussian Mixture Model.” *Advances in Neural Information Processing Systems* 12.

Siegel, David A, Menghua Wang, Stephane Maritorena, and Wayne Robinson. 2000. “Atmospheric Correction of Satellite Ocean Color Imagery: The Black Pixel Assumption.” *Applied Optics* 39 (21): 3582–91.

Van der Maaten, Laurens, and Geoffrey Hinton. 2008. “Visualizing Data Using t-SNE.” *Journal of Machine Learning Research* 9 (11).

Wyszecki, Günther, and Walter Stanley Stiles. 2000. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Vol. 40. John wiley & sons.

Zhu, Zhe, and Curtis Woodcock. 2012. “Object-Based Cloud and Cloud Shadow Detection in Landsat Imagery.” *Remote Sensing of Environment* 118 (March): 83–94. <https://doi.org/10.1016/j.rse.2011.10.028>.