Master in Green Data Science 2022/2023

Practical Machine Learning

Professor Manuel Campagnolo

Image segmentation project

Identification of Greenhouses with Satellite Images

Inês Ingham Barros da Silveira nº 22062

Joana Filipa Inácio Esteves nº 24914

Vasco de Oliveira Lourenço Florentino nº 23411

Table of Contents

Introduction 3
Data 3
Methods 4
Identification of agricultural plastic tunnels 4
Satellite Data Collection 4
Processing data 4
QGIS 4
Rasterio in Python 5
Image Segmentation 5
Hyperparameters 6
Results 6
Discussion and conclusions 7
References 9

Introduction

In recent years, the use of agricultural plastic greenhouses has significantly increased due to their ability to provide controlled environments for crop cultivation. However, the deployment of these greenhouses has led to environmental concerns, particularly regarding plastic waste management. In regions like the south of Portugal, where agricultural plastic greenhouses are prevalent, efficient identification and monitoring of these structures are essential for effective waste management and sustainable agricultural practices.

Accurate identification and segmentation of agricultural plastic greenhouses in satellite images in south of Portugal is challenging due to the complex and diverse landscape, variation in greenhouse design and size, and the presence of other similar structures like buildings and sheds. The variability in the appearance of agricultural tunnels in satellite images due to different lighting conditions, weather conditions, and seasonal changes may affect the accuracy of the algorithm, making it difficult to achieve our goal

Manual identification and mapping of these greenhouses is very time consuming, therefore, the goal of this project is to develop an image segmentation algorithm that can automatically and accurately identify agricultural plastic tunnels in satellite images. By achieving this, we aim to provide a reliable and efficient tool for agricultural and environmental experts to monitor and manage plastic waste in the agricultural sector, leading to more sustainable practices and reduced environmental impact.

Data

The data used for this work was the Sentinel-2A imagery and a shapefile with the identification of agricultural greenhouses in a specific area of the South of Portugal. All data can be found in the GitHub repository.

The satellite images were obtained from Google Earth Engine (GEE) and the shapefile containing polygons that identify the agricultural plastic greenhouses in the Mira's Irrigation Perimeter. In this case, we only used a part of the Irrigation Perimeter.

The images obtained from Google Earth Engine provided a comprehensive and updated view of the study area. These images capture the landscape in high resolution, allowing for detailed analysis and identification of various structures. The shapefile

contains polygons that outline the boundaries of the agricultural plastic tunnels. These polygons were manually identified in QGIS by Inês Silveira.

Methods

The python code for all of the steps can be found in the GitHub repository.

Identification of agricultural plastic tunnels

Performed on QGIS, the identification of the agricultural tunnels was a manual process done from February of 2022 until May of 2023. The outcome of this process was a vector format file containing polygons identifying all the greenhouses in the study area. This file is the mask used for the image segmentation project.

Satellite Data Collection

Google Earth Engine (GEE) was accessed to obtain a collection of satellite imagery, allowing us to get high-resolution images of the study area. From July of 2022, the bands downloaded were B2, B3, B4 and B8, and the coordinates for the study area are the following:

Xmin: -8.796344; Ymin: 37.440509; Xmax: -8.741399; Ymax: 37.559064

Processing data

QGIS

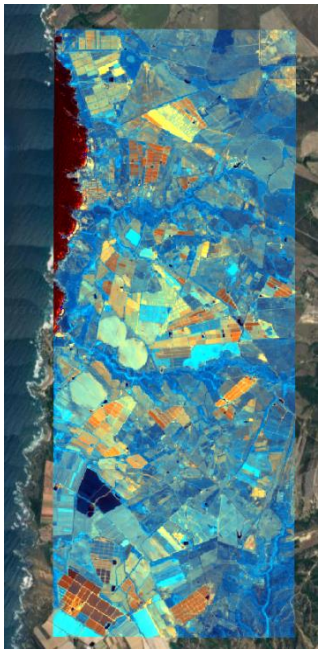
The first step was to convert the vector format mask (covered_areas_sw) into a raster file format. Using the tool "rasterize", in QGIS, the mask was created with the same resolution of the satellite imagery.

The NDVI index was calculated based on the NIR (Near-Infrared) band and red band of the visible light (see equation 1) to provide relevant information for this analysis. This helped in the visualization of the plastic greenhouses structures.

$$\text{NDVI} = \frac{(NIR-Red)}{(NIR+Red)} \quad \text{(equation 1)}$$

To create the final image for the segmentation, the bands B2, B8 and NDVI were put together with the "Build Virtual Raster" tool. The results of this steps can be seen below in figures 1, 2 and 3.





Figures 1, 2 and 3 - From left to right. Fig. 1 - polygons in red that identify the greenhouses. Fig. 2 - NDVI index. Fig. 3 - Virtual raster from NDVI, BAND 2 and BAND 8 from Sentinel-2A.

Rasterio in Python

To prepare the data for the image segmentation task, the module Rasterio was used to crop the image and its mask into smaller pieces. By extracting and working with smaller, relevant portions of the satellite images, the segmentation algorithm can achieve higher accuracy and efficiency.

Both figures were divided into 8 columns and 20 rows resulting in 160 images, based on the PRM image, and 160 images based on the mask. These were organized into "images" and "labels" into the respective folders to be used in the next step.

Image Segmentation

The segmentation technique used was U-net. This technique is a popular convolutional neural network (CNN) architecture widely used for segmentation tasks. It is known for its ability to accurately delineate object boundaries in images.

U-net is trained using a labeled dataset of input images and corresponding ground truth masks. The training dataset was 80% of the whole dataset.

Hyperparameters

The number of epochs is one of the hyperparameters that can be adjusted during the training process of a model. This refers to how many time the dataset passes through the neural network. In this case, we changed the number of epochs – try and error – to better understand which was the best-case scenario for this task to avoid under and overfitting. The model was trained with 2, 10, 20 and 30 epochs.

Results

The results presented below, represent an example of images and their predicted labels, and confusion matrices. Values of precision, recall and F1-score can be seen on table 1.

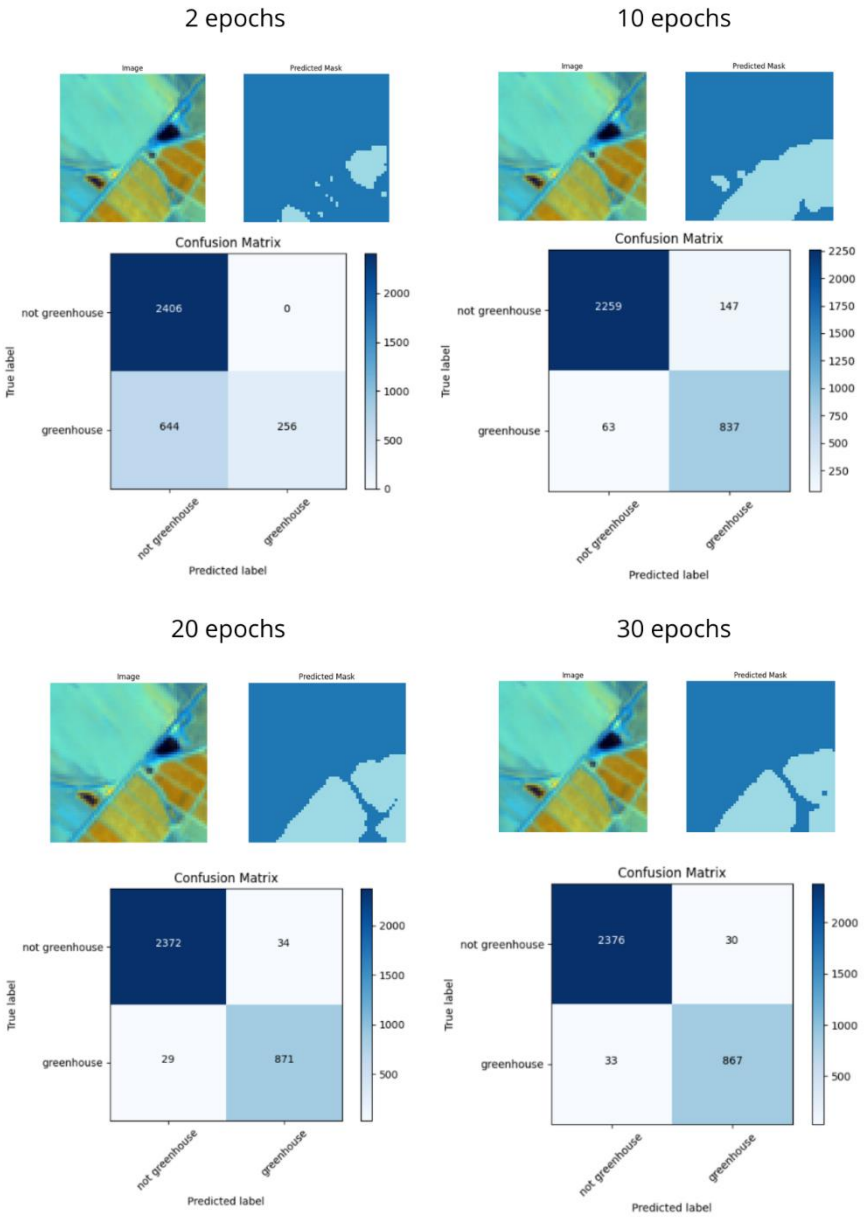
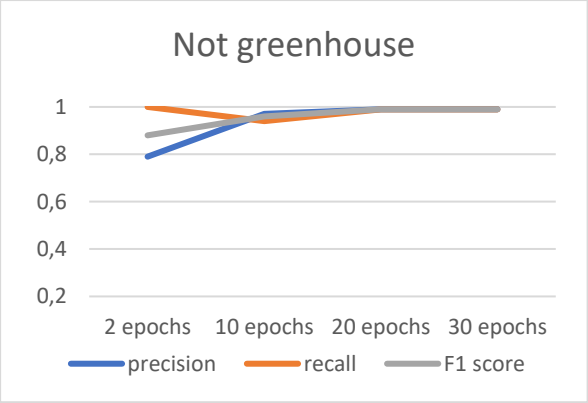


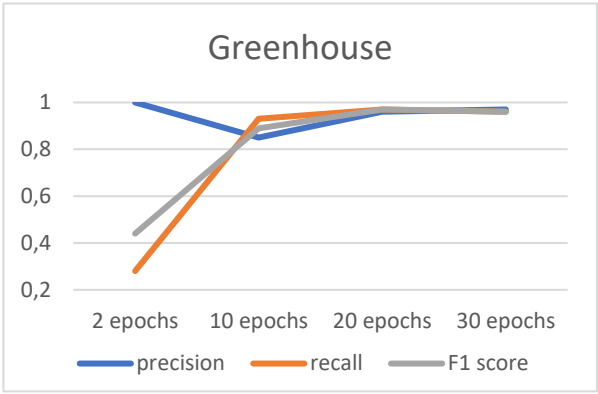
Figure 4 – Composite image, the respective predicted mask and confusion matrix for different number of epochs.

Table 1 – Values of precision, recall and F1-score for different number of epochs.

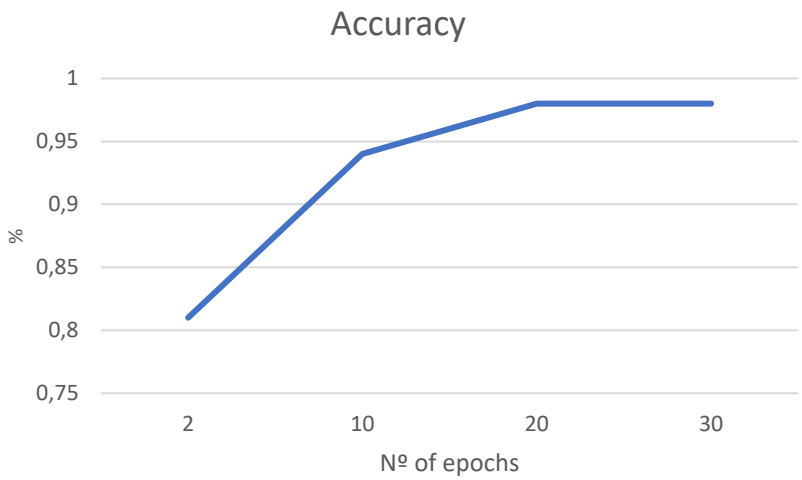
		precision	recall	f1-score	support
2 epochs	0.0	0.79	1.00	0.88	2406
	1.0	1.00	0.28	0.44	900
	accuracy			0.81	3306
	macro avg	0.89	0.64	0.66	3306
	weighted avg	0.85	0.81	0.76	3306
		precision	recall	f1-score	support
10 epochs	0.0	0.97	0.94	0.96	2406
	1.0	0.85	0.93	0.89	900
	accuracy			0.94	3306
	macro avg	0.91	0.93	0.92	3306
	weighted avg	0.94	0.94	0.94	3306
		precision	recall	f1-score	support
20 epochs	0.0	0.99	0.99	0.99	2406
	1.0	0.96	0.97	0.97	900
	accuracy			0.98	3306
	macro avg	0.98	0.98	0.98	3306
	weighted avg	0.98	0.98	0.98	3306
		precision	recall	f1-score	support
30 epochs	0.0	0.99	0.99	0.99	2406
	1.0	0.97	0.96	0.96	900
	accuracy			0.98	3306
	macro avg	0.98	0.98	0.98	3306
	weighted avg	0.98	0.98	0.98	3306



Graphic 1 - Evolution of precision, recall and F1-score for the absence of greenhouse in the predicted mask.



Graphic 2 - Evolution of precision, recall and F1-score for the presence of greenhouse in the predicted mask.



Graph 3 - Evolution of accuracy, according to the number of epochs tested.

Discussion and conclusions

Based on table 1, graphs 1, 2 and 3 were created for a general overview of the image segmentation created model.

The precision value refers to how many of the positive predictions are correct. The values for the model show a variation with the number of epochs between 0.85 to 1 for the greenhouse's classification, and 0.79 to 0.99 to not greenhouses classification.

Looking at the recall parameter, it seems to have the same pattern in both classifications. It has an opposite behavior to the precision values from 2 to 10 epochs and then follows the precision behavior for 10, 20 and 30 epochs. This parameter refers to how many of the positive cases the classifier predicted correctly, meaning that the model has a higher sensitivity when the number of epochs is higher with an exception for not greenhouse classification with 20 epochs.

Finally, F1-score combined both precision and recall and can be described as the "mean" of both metrics. That explains why the F1-score line (in grey) is represented between the precision (blue) and recall (orange) line. This metric ranges from 0 to 1, where 1 indicates the perfect precision and recall while 0 indicates the opposite, concluding the model with 20 epochs is the best fit for this task. The accuracy score (graph 3) confirms the previous statement, as the model shows a great improvement from 2 to 20 epochs, from 0.81 to 0.94, but not from 20 to 30, where is stays the same.

This indicates that this model has not reached an optimal parameterization and more experimentation with parameters should be made, with more intermediate values, to better understand the slopes of functions of the values.

There are few other adjustments that can be made to improve the model, such as changing the number of channels and parameters, and changing the split of the training dataset. There could also be a comparison of the model with the testing and validation dataset and respective error values and cross-validation score.

Concluding, this model can predict with an accuracy of 98% the greenhouses, with 20 epochs, being this a good enough value to trust for large scale estimates in Portugal and similar conditions.

References

- Measuring vegetation (NDVI & EVI). (2000, August 30). Available at: https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php, accessed in 6th of June of 2023

- https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec#:~:text=F1%2Dscore%20equals%20precision%20and%20recall%20if%20the%20two%20input,examples%20they%20start%20to%20vary, accessed on 12th of June of 2023

- Pereira Cardoso, P.M (2022). Monitoring Greenhouses with Satellite Images and Machine Learning. Engenharia de Redes e Sistemas Informáticos. Departamento de Ciência dos Computadores, Faculdade de Ciências da Universidade do Porto.