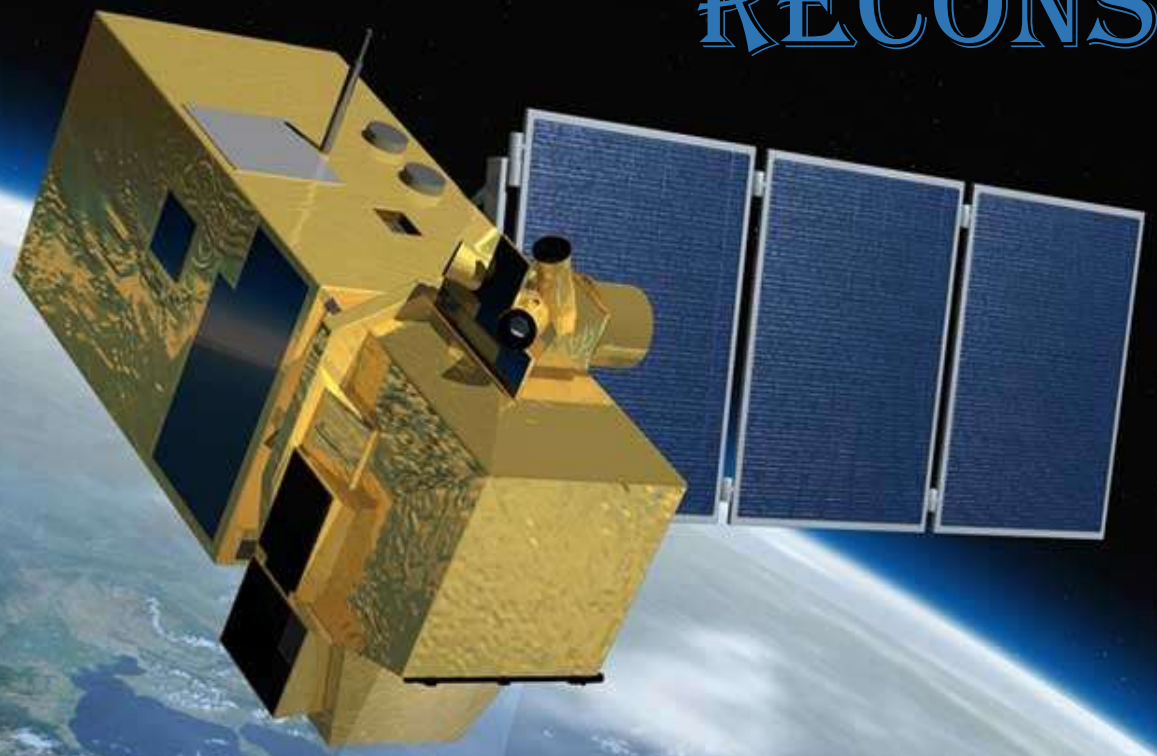
FLAGSHIP: AI ALGORITHM FOR (SATELLITE) IMAGING RECONSTRUCTION



REPORT FOR WP6 MEETING, 11/10/2023

Alessia Tricomi[1,2,3], Giuseppe Piparo[1,2]

1. INFN Sezione di Catania
2. Università degli studi di Catania
3. Centro Siciliano di Fisica Nucleare e Struttura della Materia (CSFNSM)

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

STATUS OF ACTIVITIES

- We are currently in an '**exploratory**' phase of the possibilities and the challenges offered by satellite data analysis.

- Understanding data, following the **'3V rule'**: What Volume of data do we have? How Varied are the types of information? Velocity of acquisition?

- What **difficulties** can be encountered when analysing satellite data? What are the most important **limiting factors**? How can we go beyond **current knowledge**?

- Can the use of **machine learning techniques** be decisive for image analysis? Which **architectures** are preferable? How can we build high-quality datasets to train, validate and test models?

PRELIMINARY TESTS

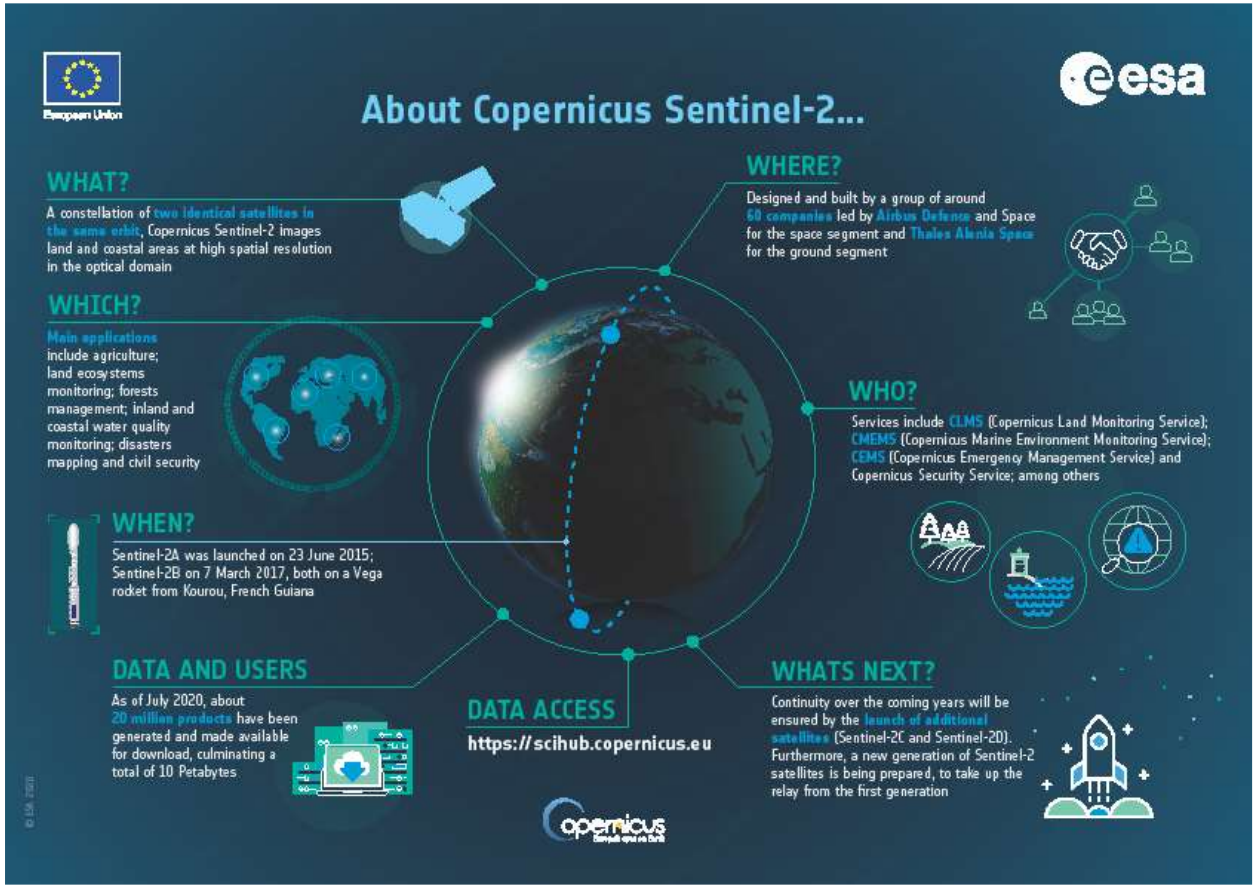
- **We ran the first tests to start finding answers to the questions on the previous slide.**
- First of all, we started by understanding how to access satellite data in a **simple, organic and scalable way**.
- We may have found a possibility using the python library SentinelHub (site: Sentinel Hub (sentinel-hub.com), git repo: GitHub - sentinel-hub/sentinelhub-py: Download and process satellite imagery in Python using Sentinel Hub services.).
- **SentinelHub makes it possible to download data from the Copernicus Sentinel satellites.** We are currently using the 'basic' version and only accessing public data.

SENTINEL-2

- For testing we are currently only using data from **Sentinel-2A** (tests on other satellites will follow).

SENTINEL-2

- For testing we are currently only using data from **Sentinel-2A** (tests on other satellites will follow).
- Sentinel-2 is specifically designed for **vegetation monitoring** and **managing natural disasters.**



SENTINEL-2

- For testing we are currently only using data from **Sentinel-2A** (tests on other satellites will follow).
- Sentinel-2 is specifically designed for **vegetation monitoring** and **managing natural disasters.**
- It takes data at **different wavelengths**, which can be composed into indices with strong discriminating power **(e.g. NDVI, NDMI, etc.)**.

Name	Description	Resolution
B01	Coastal aerosol, 442.7 nm (S2A), 442.3 nm (S2B)	60m
B02	Blue, 492.4 nm (S2A), 492.1 nm (S2B)	10m
B03	Green, 559.8 nm (S2A), 559.0 nm (S2B)	10m
B04	Red, 664.6 nm (S2A), 665.0 nm (S2B)	10m
B05	Vegetation red edge, 704.1 nm (S2A), 703.8 nm (S2B)	20m
B06	Vegetation red edge, 740.5 nm (S2A), 739.1 nm (S2B)	20m
B07	Vegetation red edge, 782.8 nm (S2A), 779.7 nm (S2B)	20m
B08	NIR, 832.8 nm (S2A), 833.0 nm (S2B)	10m
B8A	Narrow NIR, 864.7 nm (S2A), 864.0 nm (S2B)	20m
B09	Water vapour, 945.1 nm (S2A), 943.2 nm (S2B)	60m
B11	SWIR, 1613.7 nm (S2A), 1610.4 nm (S2B)	20m
B12	SWIR, 2202.4 nm (S2A), 2185.7 nm (S2B)	20m
AOT	Aerosol Optical Thickness map, based on Sen2Cor processor	10m
SCL	Scene classification data, based on Sen2Cor processor, codelist	20m
SNW	Snow probability, based on Sen2Cor processor	20m
CLD	Cloud probability, based on Sen2Cor processor	20m
CLP	Cloud probability, based on s2cloudless (more)	160m
CLM	Cloud masks (more)	160m
sunAzimuthAngles	Sun azimuth angle	5000m
sunZenithAngles	Sun zenith angle	5000m
viewAzimuthMean	Viewing azimuth angle	5000m
viewZenithMean	Viewing zenith angle	5000m
dataMask	The mask of data/no data pixels (more)	N/A*

SENTINEL-2

- For testing we are currently only using data from **Sentinel-2A** (tests on other satellites will follow).
- Sentinel-2 is specifically designed for **vegetation monitoring** and **managing natural disasters.**
- It takes data at **different wavelengths**, which can be composed into indices with strong discriminating power **(e.g. NDVI, NDMI, etc.)**.
- Being able to use several bands in many combinations, the Sentinel-2 data lends itself well to **multivariate analyses**.

Name	Description	Resolution
B01	Coastal aerosol, 442.7 nm (S2A), 442.3 nm (S2B)	60m
B02	Blue, 492.4 nm (S2A), 492.1 nm (S2B)	10m
B03	Green, 559.8 nm (S2A), 559.0 nm (S2B)	10m
B04	Red, 664.6 nm (S2A), 665.0 nm (S2B)	10m
B05	Vegetation red edge, 704.1 nm (S2A), 703.8 nm (S2B)	20m
B06	Vegetation red edge, 740.5 nm (S2A), 739.1 nm (S2B)	20m
B07	Vegetation red edge, 782.8 nm (S2A), 779.7 nm (S2B)	20m
B08	NIR, 832.8 nm (S2A), 833.0 nm (S2B)	10m
B8A	Narrow NIR, 864.7 nm (S2A), 864.0 nm (S2B)	20m
B09	Water vapour, 945.1 nm (S2A), 943.2 nm (S2B)	60m
B11	SWIR, 1613.7 nm (S2A), 1610.4 nm (S2B)	20m
B12	SWIR, 2202.4 nm (S2A), 2185.7 nm (S2B)	20m
AOT	Aerosol Optical Thickness map, based on Sen2Cor processor	10m
SCL	Scene classification data, based on Sen2Cor processor, codelist	20m
SNW	Snow probability, based on Sen2Cor processor	20m
CLD	Cloud probability, based on Sen2Cor processor	20m
CLP	Cloud probability, based on s2cloudless (more)	160m
CLM	Cloud masks (more)	160m
sunAzimuthAngles	Sun azimuth angle	5000m
sunZenithAngles	Sun zenith angle	5000m
viewAzimuthMean	Viewing azimuth angle	5000m
viewZenithMean	Viewing zenith angle	5000m
dataMask	The mask of data/no data pixels (more)	N/A*

SOME SENTINEL-2 IMAGES

- Region near **Simeto** river (Catania), using RGB bands (B02, B03, B04). Very good resolution (10 m). But this is a good image…

SOME SENTINEL-2 IMAGES

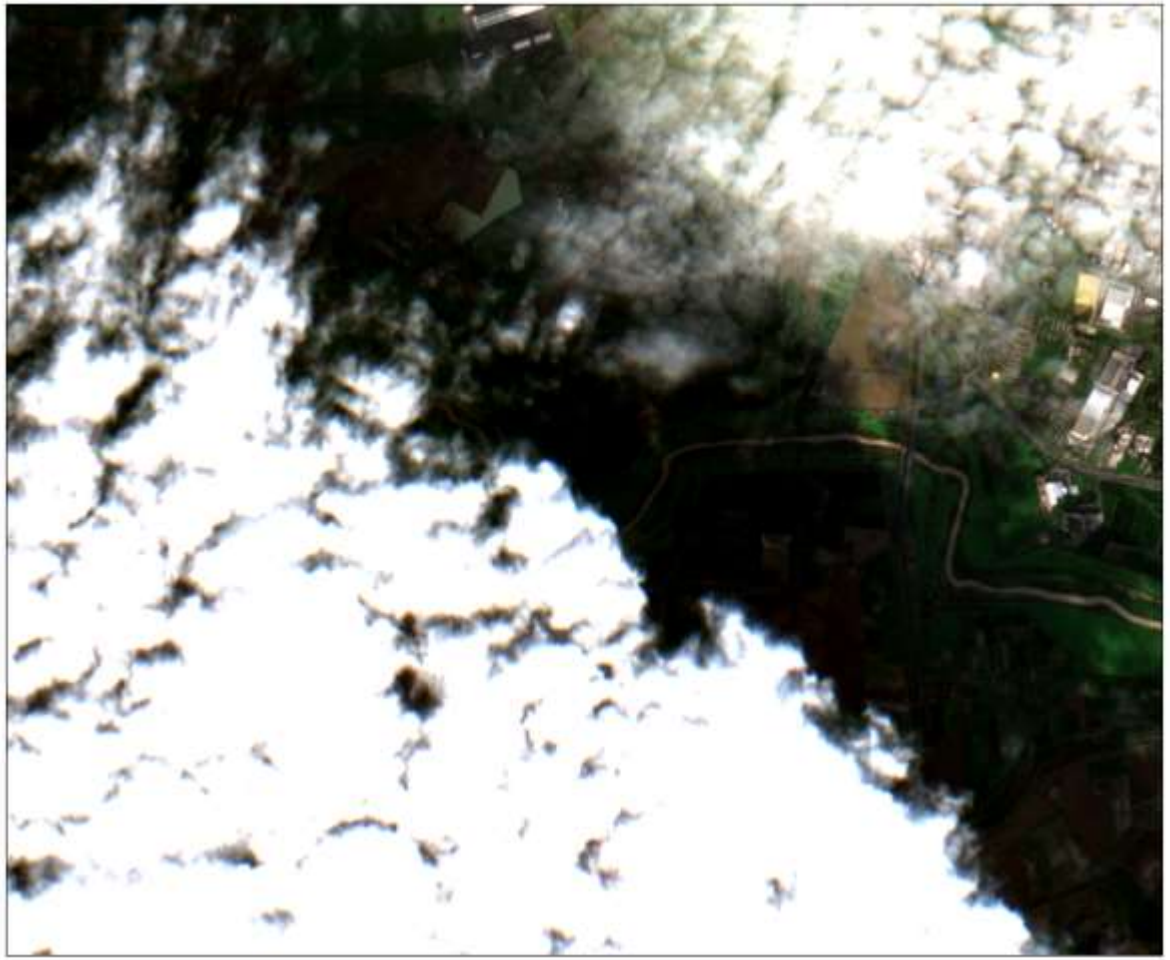
- Region near **Simeto** river (Catania), using RGB bands (B02, B03, B04). Very good resolution (10 m). But this is a good image…

- **This one is a little less better...**

SOME SENTINEL-2 IMAGES

- Region near **Simeto** river (Catania), using RGB bands (B02, B03, B04). Very good resolution (10 m). But this is a good image…

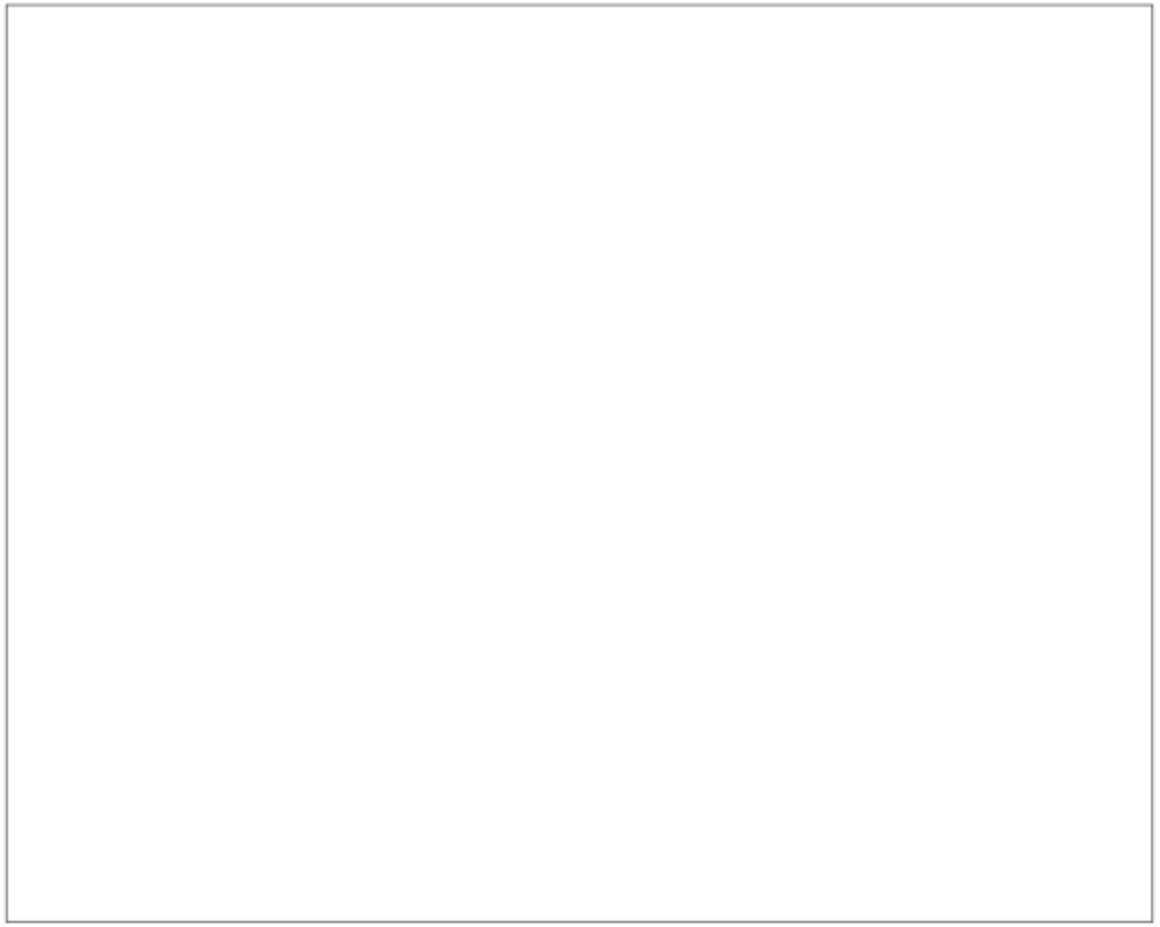
- **This one is a little less better...**
- **This one is a lot less better...**



SOME SENTINEL-2 IMAGES

- Region near **Simeto** river (Catania), using RGB bands (B02, B03, B04). Very good resolution (10 m). But this is a good image…

- **This one is a little less better...**
- **This one is a lot less better...**
- **This is unusable!**

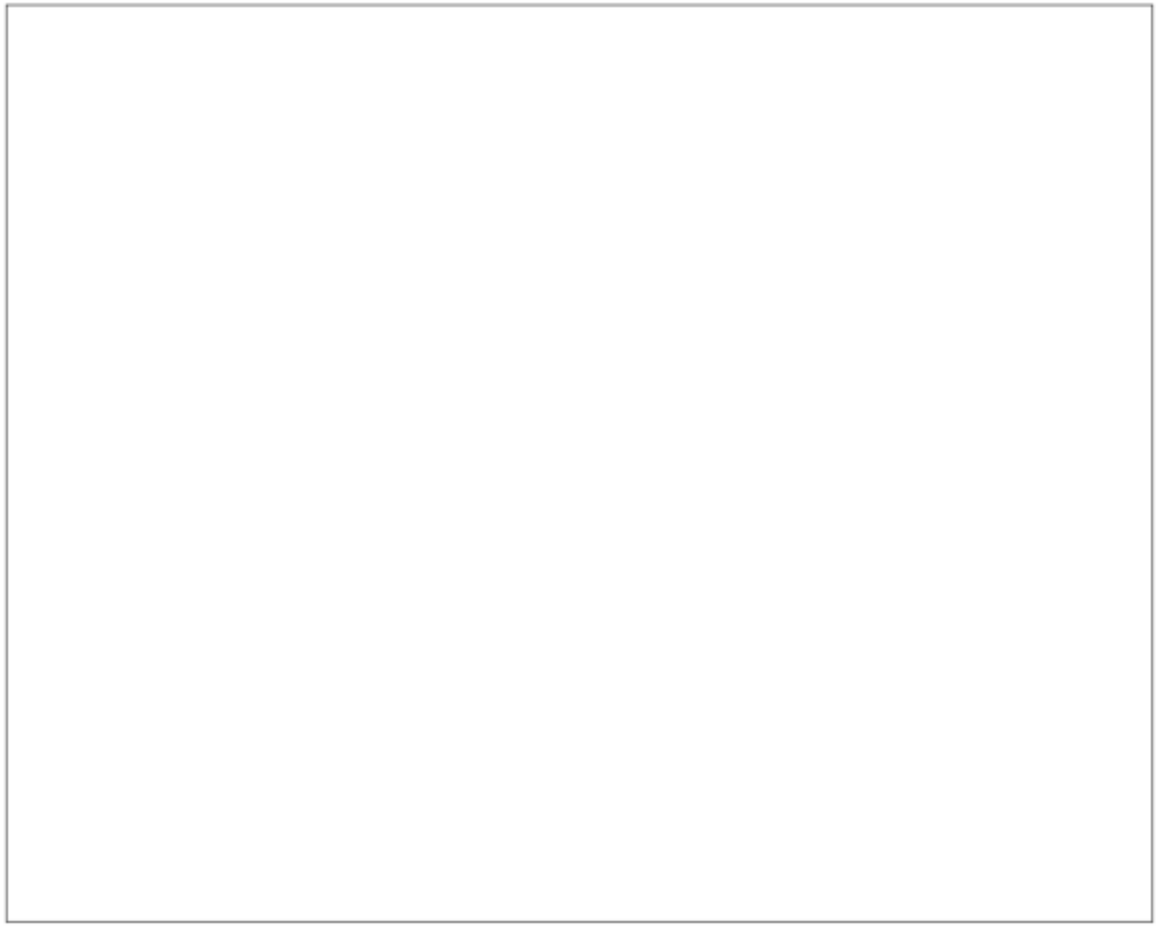


SOME SENTINEL-2 IMAGES

- Region near **Simeto** river (Catania), using RGB bands (B02, B03, B04). Very good resolution (10 m). But this is a good image…

- **This one is a little less better...**
- **This one is a lot less better...**
- **This is unusable!**

- Taking cloudiness into account is **VERY IMPORTANT**, especially in ML algorithms. It could be a bias in the training/testing phase. Or it could be the turning factor. Depends on the applications…



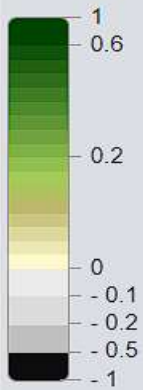
LET'S TRY A ML APPLICATION!

- After understanding how to download, save and visualize data in Sentinel-2, we tried to develop some ML algorithms. But immediately the first problem arises...
- Let's say we want to analyse the Simeto region shown earlier and understand the health of the vegetation. To use a supervised algorithm, **we need a (well) labelled dataset**. Without knowledge or access to specific information, this is really difficult!
- Unsupervised algorithms could be used (a **PCA+Clusterer** approach is currently under study, but preliminary results are not given in this presentation), **but even then, the interpretation of the results would have to be performed by an expert!**

In order to at least begin to carry out some tests, we have realized a rather trivial, but in its own way interesting algorithm. A convolutional network capable of identifying the season in the analyzed region (trivial labelling problem and easy-to-test results).

SEASON RECOGNITION DEVELOPMENT PIPELINE

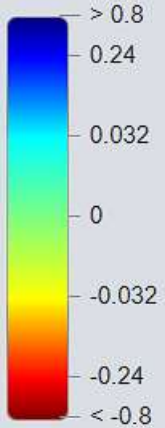
1. **Select the data format:** We used images of the region in the four seasons from 2018 to 2022, with a frequency of about 5 days (**i.e. the Sentinel-2A acquisition rate**). We used two indices and one band, namely NDVI=(B08-B04)/(B08+B04), Moisture Index=(B8A - B11)/(B8A+B11) and SWIR=B12.



Normalized Difference Vegetation Index (NDVI)

The normalized difference vegetation index is a simple, but effective index for quantifying green vegetation. It is a measure of the state of vegetation health based on how plants reflect light at certain wavelengths. The value range of the NDVI is -1 to 1. Negative values of NDVI (values approaching -1) correspond to water. Values close to zero (-0.1to 0.1) generally correspond to barren areas of rock, sand, or snow. Low, positive values represent shrub and grassland (approximately 0.2 to 0.4), while high values indicate temperate and tropical rainforests (values approaching 1).

More info here and here.



Normalized Difference Moisture Index (NDMI)

The normalized difference moisture Index (NDMI) is used to determine vegetation water content and monitor droughts. The value range of the NDMI is -1 to 1. Negative values of NDMI (values approaching -1) correspond to barren soil. Values around zero (-0.2 to 0.4) generally correspond to water stress. High, positive values represent high canopy without water stress (approximately 0.4 to 1).

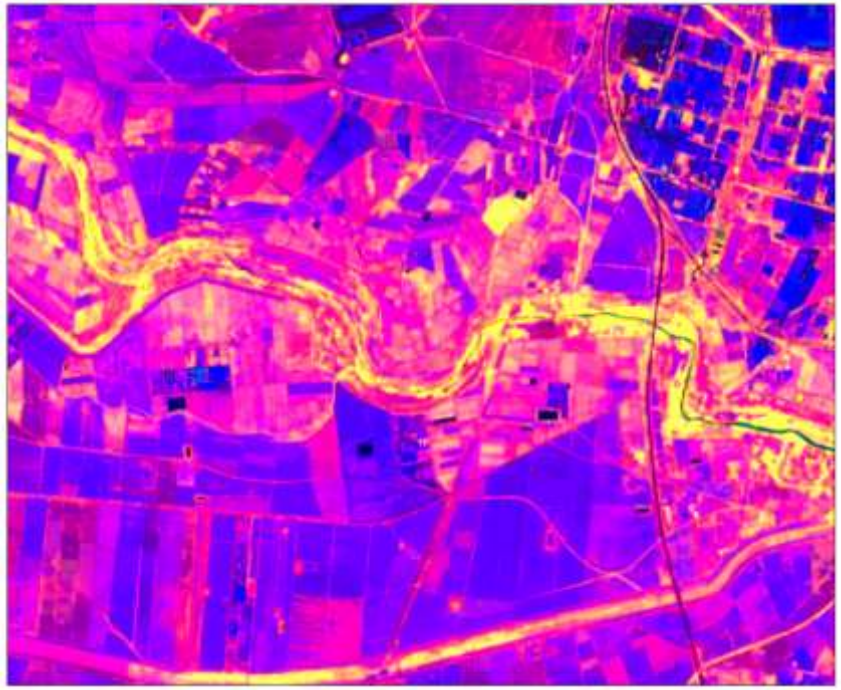
More info here.

Short wave infrared composite (SWIR)

Short wave infrared (SWIR) measurements can help scientists estimate how much water is present in plants and soil, as water absorbs SWIR wavelengths. Short wave infrared bands (a band is a region of the electromagnetic spectrum; a satellite sensor can image Earth in different bands) are also useful for distinguishing between cloud types (water clouds versus ice clouds), snow and ice, all of which appear white in visible light. In this composite vegetation appears in shades of green, soils and built-up areas are in various shades of brown, and water appears black. Newly burned land reflects strongly in SWIR bands, making them valuable for mapping fire damages. Each rock type reflects shortwave infrared light differently, making it possible to map out geology by comparing reflected SWIR light.

SEASON RECOGNITION DEVELOPMENT PIPELINE

1. **Select the data format:** We used images of the region in the four seasons from 2018 to 2022, with a frequency of about 5 days (**i.e. the Sentinel-2A acquisition rate**). We used two indices and one band, namely NDVI=(B08-B04)/(B08+B04), Moisture Index=(B8A - B11)/(B8A+B11) and SWIR=B12.



SUMMER

Combination of 3 indices/band

WINTER

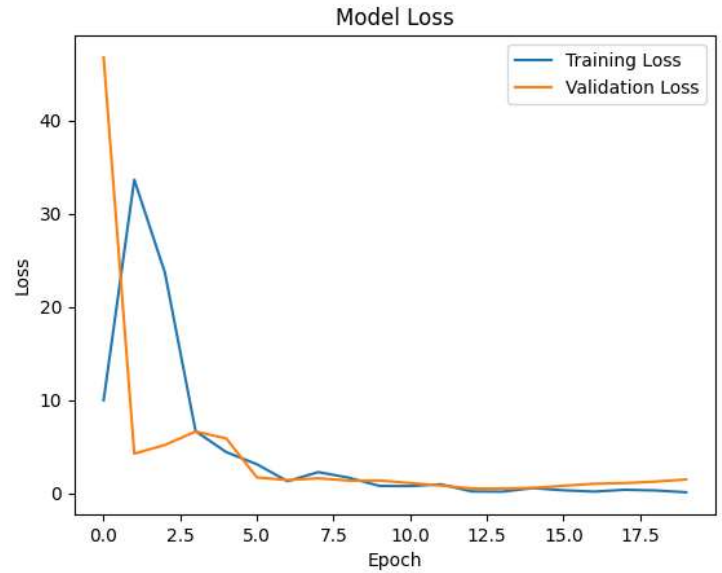
SEASON RECOGNITION DEVELOPMENT PIPELINE

1. **Select the data format:** We used images of the region in the four seasons from 2018 to 2022, with a frequency of about 5 days (**i.e. the Sentinel-2A acquisition rate**). We used two indices and one band, namely NDVI=(B08-B04)/(B08+B04), Moisture Index=(B8A - B11)/(B8A+B11) and SWIR=B12.
2. **Build a ML model for season recognition:** We have chosen a trivial **Convolution Neural Network (CNN)** for the first tests. The fine-tuning of the hyperparameters is in progress, but the network doesn't seem to work badly or over/under-fit too much.

```
model = Sequential([
    Conv2D(32, (3, 3), activation='relu', input_shape=(615, 747, 3)),  #kernel_regularizer=l2(regularization_strength)
    MaxPooling2D(2, 2),
   # Dropout(0.25),
    Conv2D(64, (3, 3), activation='relu'),
    MaxPooling2D(2, 2),
    #Dropout(0.25),
    Flatten(),
    Dense(128, activation='relu', ),
    Dropout(0.5),  # Dropout layer
    Dense(4, activation='softmax')
])

model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

history = model.fit(np.array(X_train), np.array(y_train), epochs=20, validation_data=(np.array(X_test), np.array(y_test)),callbacks=[early_stopping])
```

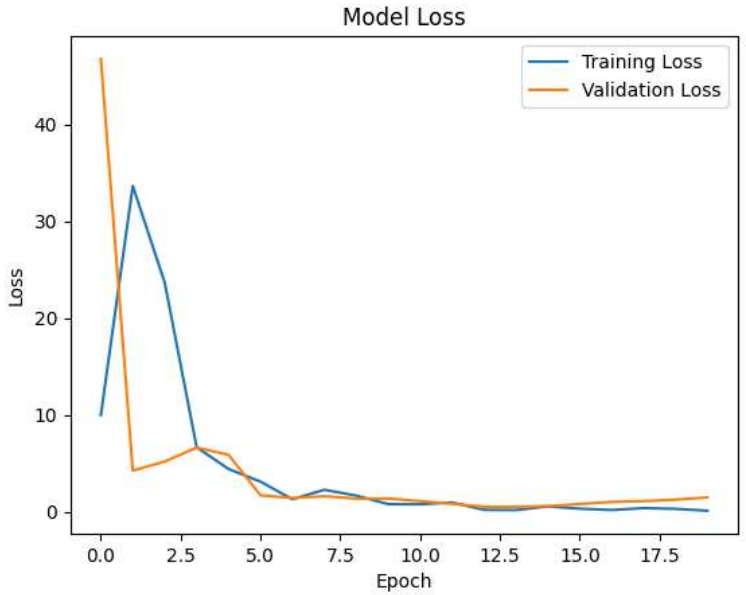


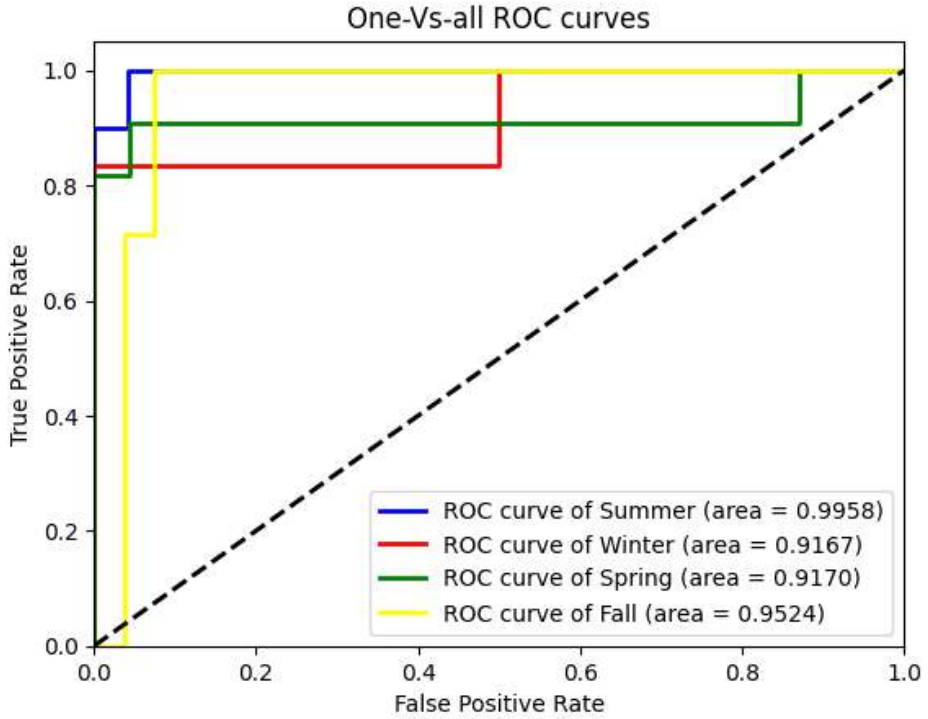
SEASON RECOGNITION DEVELOPMENT PIPELINE

1. **Select the data format:** We used images of the region in the four seasons from 2018 to 2022, with a frequency of about 5 days (**i.e. the Sentinel-2A acquisition rate**). We used two indices and one band, namely NDVI=(B08-B04)/(B08+B04), Moisture Index=(B8A - B11)/(B8A+B11) and SWIR=B12.
2. **Build a ML model for season recognition:** We have chosen a trivial **Convolution Neural Network (CNN)** for the first tests. The fine-tuning of the hyperparameters is in progress, but the network seems It doesn't seem to work badly or over/under-fit too much.
3. **Results: The model is able to recognize the season with a nice efficiency, even if low statistics for training/test are provided (due to the low acquisition rate of Sentinel-2A), some "bad" figures are present (cloud even if filtered), we didn't train it too much (only 20 epochs) and we didn't fine-tune well the hyperparameters. In addition, the ML model is very trivial and "homemade". We could consider the possibility of using high-level models with transfer learning for real use-case**

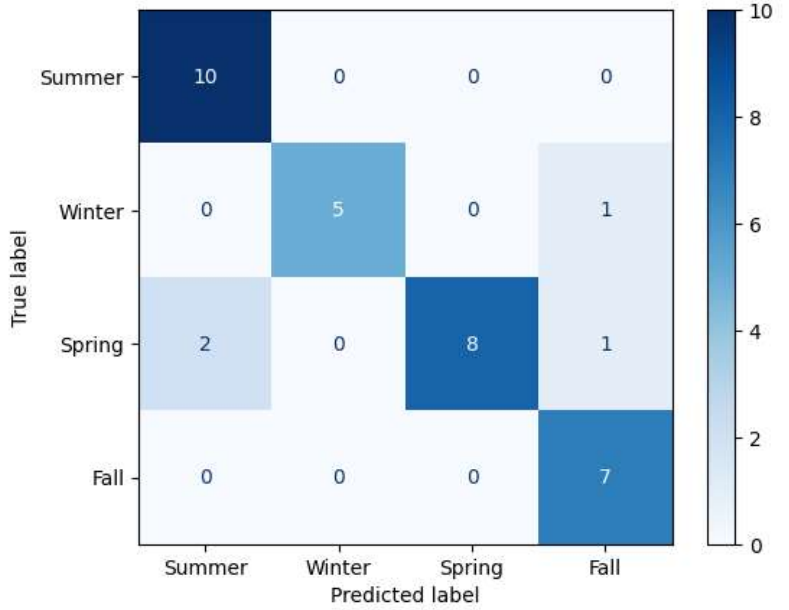
SEASON RECOGNITION DEVELOPMENT PIPELINE

Results: **The model is able to recognize the season with a nice efficiency, even if low statistics for training/test are provided (due to the low acquisition rate of Sentinel-2A), some "bad" figures are present (cloud even if filtered), we didn't train it too much (only 20 epochs) and we didn't fine-tune well the hyperparameters. In addition, the ML model is very trivial and "homemade". We could consider the possibility of using high-level models with transfer learning for real use-case**





Confusion matrix



What we learned

- **At the end of this simple application, we learned several important things:**
  1. **Copernicus satellite datasets are very promising!** It could offer a large Variety and Volume of information, that can be handled using ML techniques.
  2. Velocity i.e. acquisition rate is good, but some short-time information could be lost… **Must be considered in the analysis**.
  3. **Also, clouds could be problematic to be treated**. Luckily, many techniques and specifically designed information of other satellites are available (see SAR observation).
  4. **CNN are valid models to treat with this kind of data**. The use of **transfer learning** could be very promising.
  5. A great initial difficulty is to catalogue datasets. Not even the most advanced model ever created could perform well with bad data.
  6. **Unsupervised models** also deserve consideration for these applications, but their interpretation may be difficult. (**See next time**.)

MILESTONES

1. M1-M6 (corresponding to MS7): Survey of the State-of-the-Art; tracking of R&D technologies to be used; selection of datasets for use cases (at least one).
   - ➢ D1: report on technologies to be used, selection of at least one test dataset.
2. M7-M10 (corresponding to MS8): first experimentation with data sources and algorithms, demonstration on the feasibility of choices
   - ➢ D2: report on the experimentation and of technical choices; first code repository available
3. M11-M24 (corresponding to MS10): Implementation of the selected technology(ies); test and validation on selected dataset(s). Proof-of-Concept deployment.
   - ➢ D3: Report on the work carried out; release of the developed code on public repository.
   - ➢ Intermediate report at MS9