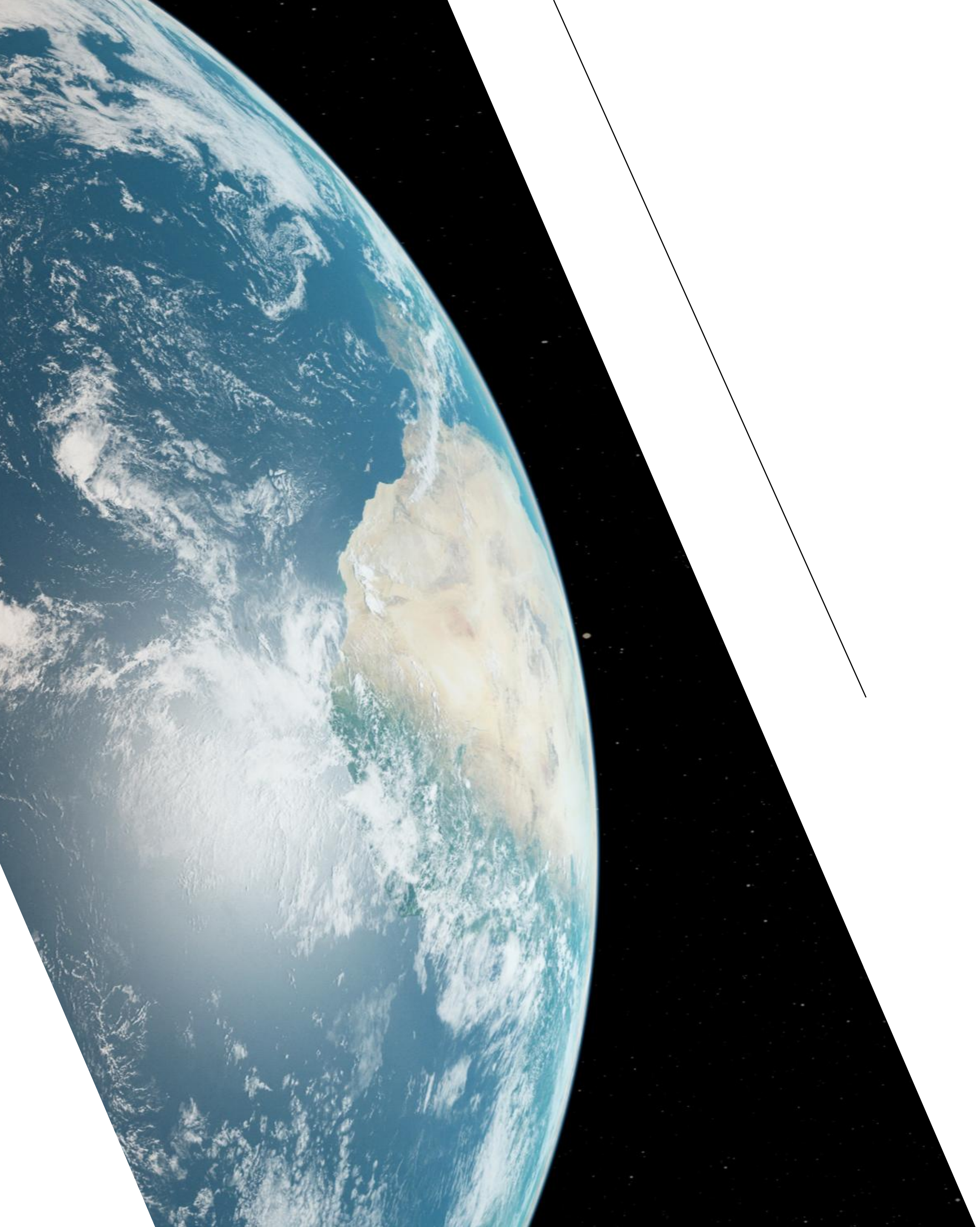
ARSS PROJECT

REIMPLEMENTATION OF A LAND COVER CLASSIFICATION PIPELINE

PROJECT FOR «*SENSING AND RADAR TECHNOLOGIES*» AND «*PROJECT COURSE*»



SUMMARY

1. Introduction;
2. Pipeline Structure;
3. First component: **Masks Refinement**;
4. Second component: **Cloud Restoration**;
5. Third component: **Features Generation**;
6. Fourth component: **Dataset Generation**;
7. Fifth component: **SVM Pipeline**;
8. Sixth component: **Land Cover Maps Generation**.

INTRODUCTION

The primary goal was to reimplement the land cover classification pipeline developed under the ESA Climate Change Initiative's High Resolution Land Cover (CCI HRLC) project, and to conduct experiments on various components to assess their impact on the final classification results.

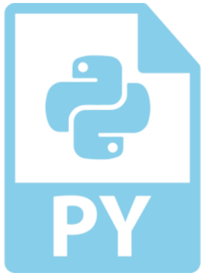
This pipeline utilizes Sentinel-2 Level-2A (L2A) products to generate 10-meter resolution land cover maps over selected regions, supporting climate monitoring and land analysis.

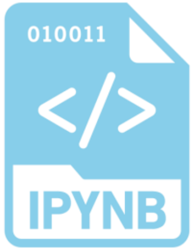
Reimplementation strategy

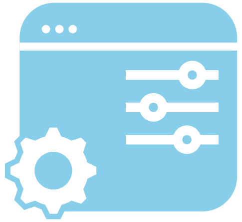
- Use xarray and dask to improve efficiency and scalability;
- Evaluate mask quality using CloudSEN12+ reference data;
- Test new features and training strategies to assess impact on LC maps;
- Compare performance with the existing pipeline;
- Build a more robust and modular baseline, reusing core logic from the original workflow.

Project structure

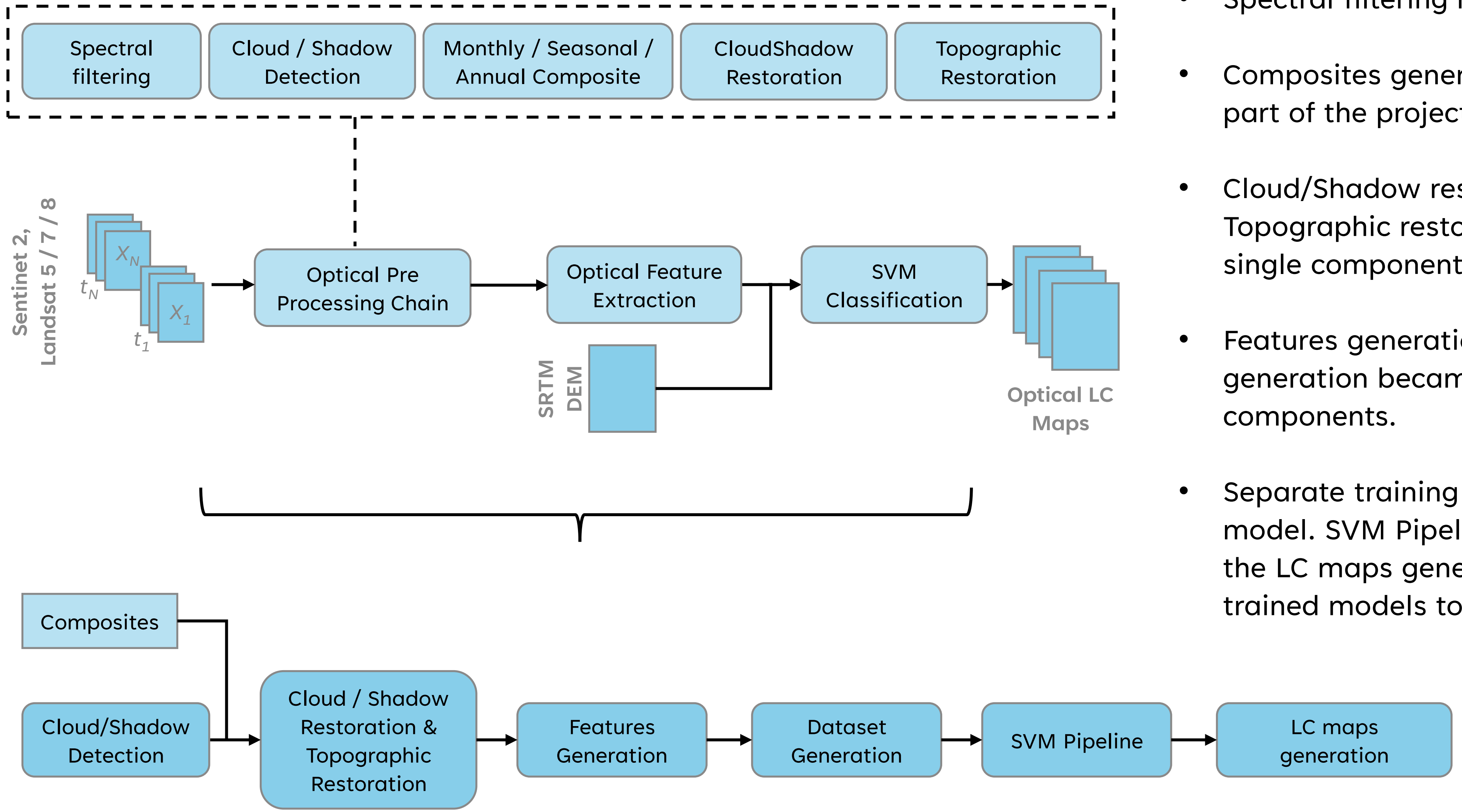
- Developed entirely within a Docker container for consistent execution across environments;
- Each pipeline component includes:
    - A modular Python script implementing the full logic;
    - An interactive Jupyter notebook for step-by-step testing;
    - A .yaml file defining all execution parameters.



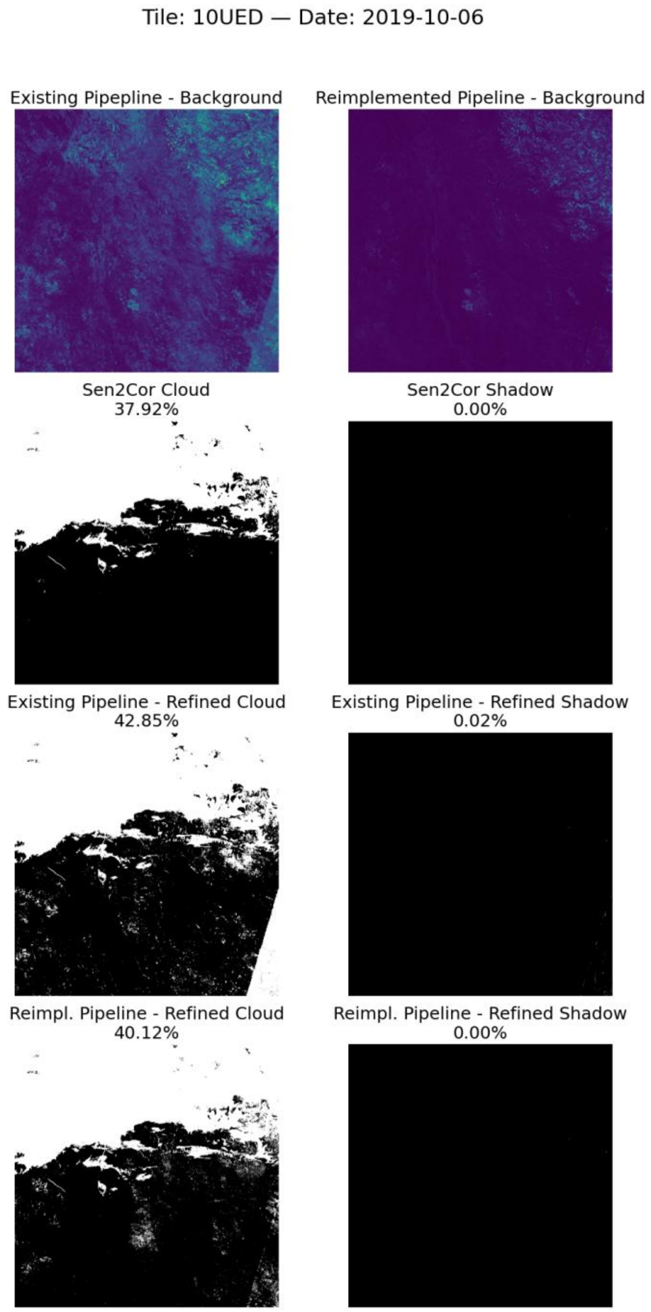




PIPELINE STRUCTURE



- Spectral filtering removed.
- Composites generated externally and not part of the project.
- Cloud/Shadow restoration and Topographic restoration integrated in a single component.
- Features generation and dataset generation became two separate components.
- Separate training and usage of the SVM model. SVM Pipeline trains the SVM and the LC maps generation steps uses trained models to classify entire tiles.



FIRST COMPONENT: MASK REFINEMENT

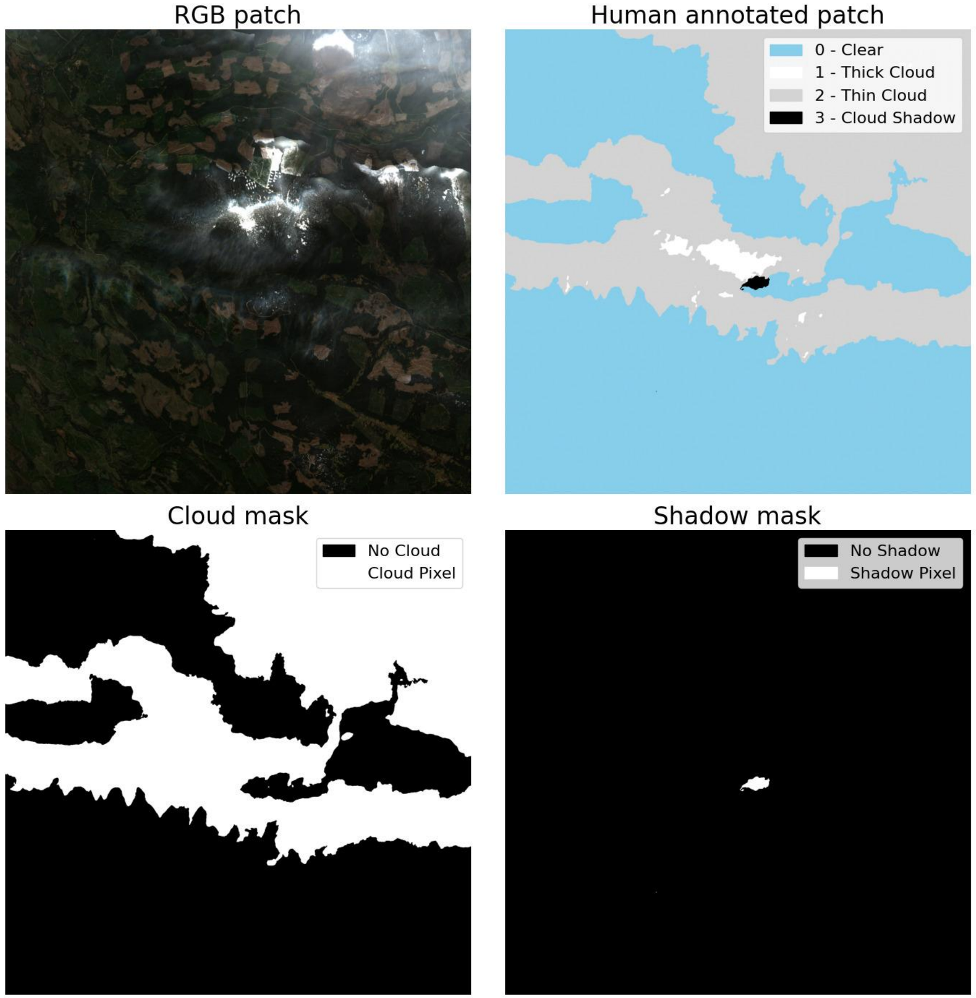
The first component of the pipeline refines cloud and shadow masks from Sentinel-2 L2A products, leveraging the Scene Classification Layer (SCL) and adapting the original logic to work with xarray.

Main steps:

- Load L2A products as lazy xarray datasets
- Extract initial Sen2Cor cloud and shadow masks from the SCL
- Generate seasonal background images using a quantile-based aggregation of the blue band
- Apply K-means clustering to refine cloud masks based on deviations from the background
- Detect cloud shadows using the Cloud Shadow Index (CSI) and blue reflectance thresholds
- Save refined masks for downstream processing

MASKS REFINEMENT – EVALUATION WITH

The refined cloud and shadow masks were evaluated using the CloudSEN12+ dataset, which provides high-quality, human-annotated samples. Only 2000 × 2000 - pixel patches from tile ***18NWL (2019-02-23)*** were used in the analysis presented below.



Cloud masks evaluation

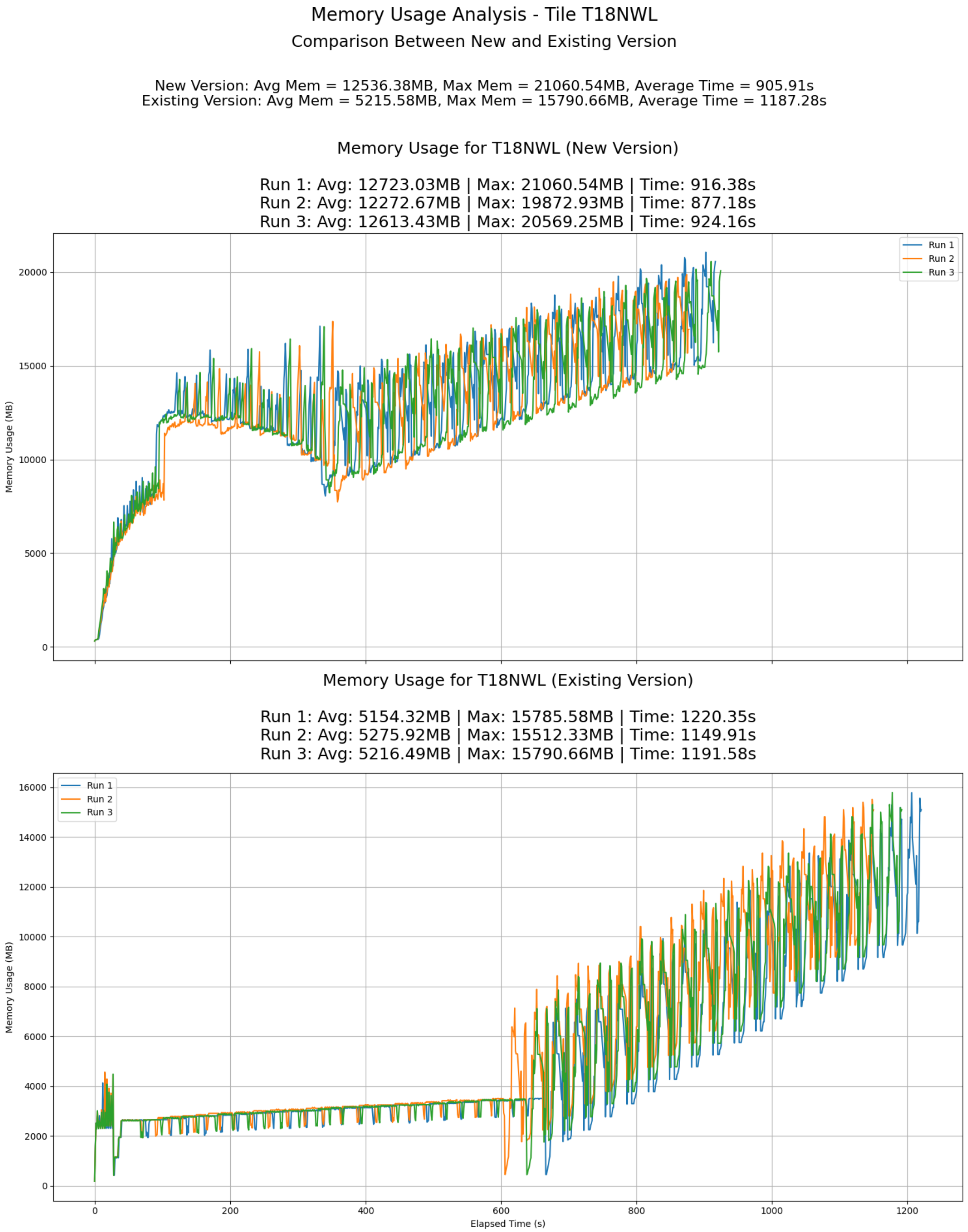
- Improved precision and F1-Score with the reimplmeneted pipeline
- Slightly lower recall compared to existing pipeline

Pipeline	Mask Type	Precision	Recall	F1 - Score
Existing	Sen2Cor Cloud Refined Cloud	0.55 0.51	0.89 0.90	0.68 0.65
Reimplemented	Sen2Cor Cloud Refined Cloud	0.55 0.55	0.89 0.89	0.68 0.68

Shadow masks evaluation

- Improved F1-Score and Recall with reimplmented pipeline
- Lower precision

Pipeline	Mask Type	Precision	Recall	F1 - Score
Existing	Sen2Cor Shadow Refined Shadow	0.76 0.76	0.07 0.07	0.13 0.13
Reimplemented	Sen2Cor Shadow Refined Shadow	0.76 0.72	0.07 0.08	0.13 0.14



MASKS EVALUATION: PERFORMANCE

The performance of the reimplemented component was evaluated against the same component of the previous pipeline. The testing configuration was as follow:

- Both the component were executed 3 times to have more consistent results;
- The evaluation involve the memory used and the time taken by each execution;
- New pipeline achieves ~25% faster execution time at the cost of higher memory usage (avg. 12.5 GB vs. 5.2 GB).

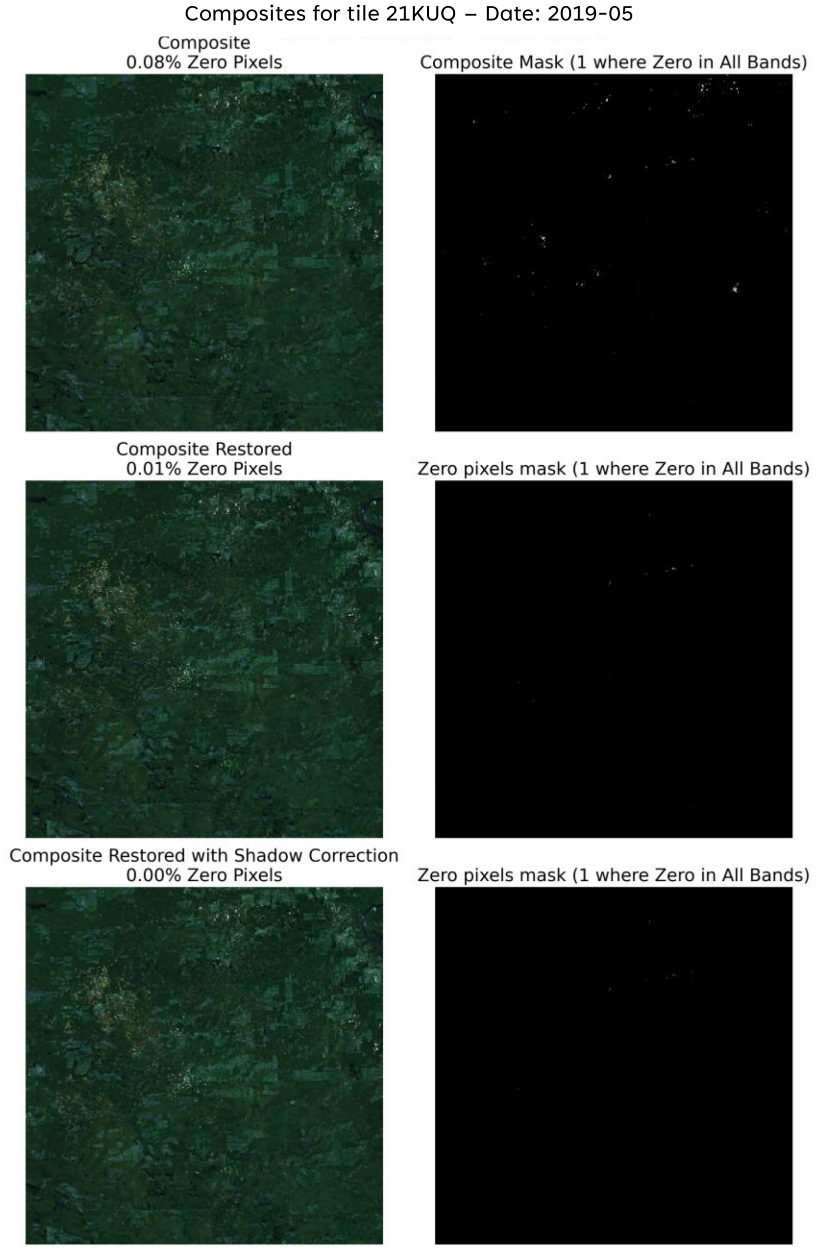
SECOND COMPONENT: CLOUD RESTORATION

This component restores missing pixels in monthly Sentinel-2 composites, which are generated externally using all available L2A products for each month. Although composite generation is not part of this project, the restoration process is fully implemented.
The adopted strategy replicates the logic of the original pipeline and consists of two sequential steps:

- **Temporal Interpolation:** for each pixel, if the value is missing in a given month, the same pixel is retrieved from the adjacent monthly composites. If both previous and next values are available, the missing pixel is filled using their mean.

- **Topographic Correction:** residual shadowed areas caused by terrain are corrected using slope information derived from the Digital Elevation Model (DEM). The restored composite is adjusted based on statistical relationships between shadowed and non-shadowed regions.

- **Main Difference:** while the core logic is preserved from the original implementation, the reimplemented version is fully based on xarray, which allows lazy evaluation and chunk-wise processing, significantly improving scalability and memory efficiency.



THIRD COMPONENT: FEATURES GENERATION

This component extracts the full set of features required for land cover classification, using Sentinel-2 monthly composites and topographic information from DEM data. It reimplements and extends the logic of the original pipeline in a modular and reproducible form.

The generated features include:

- ***6 texture features*** based on Gray-Level Co-occurrence Matrix (GLCM), consistent with the original approach
- ***3 topographic features**:* elevation, slope, and aspect
- ***144 spectral features*** from composite bands
- ***14 additional features*** introduced for experimental purposes. These new features are designed to capture a broader range of land surface characteristics, including vegetation health, moisture content, bare soil exposure, built-up structures, and landscape boundaries. Their inclusion allows investigation into whether richer descriptors improve class separability in land cover mapping.

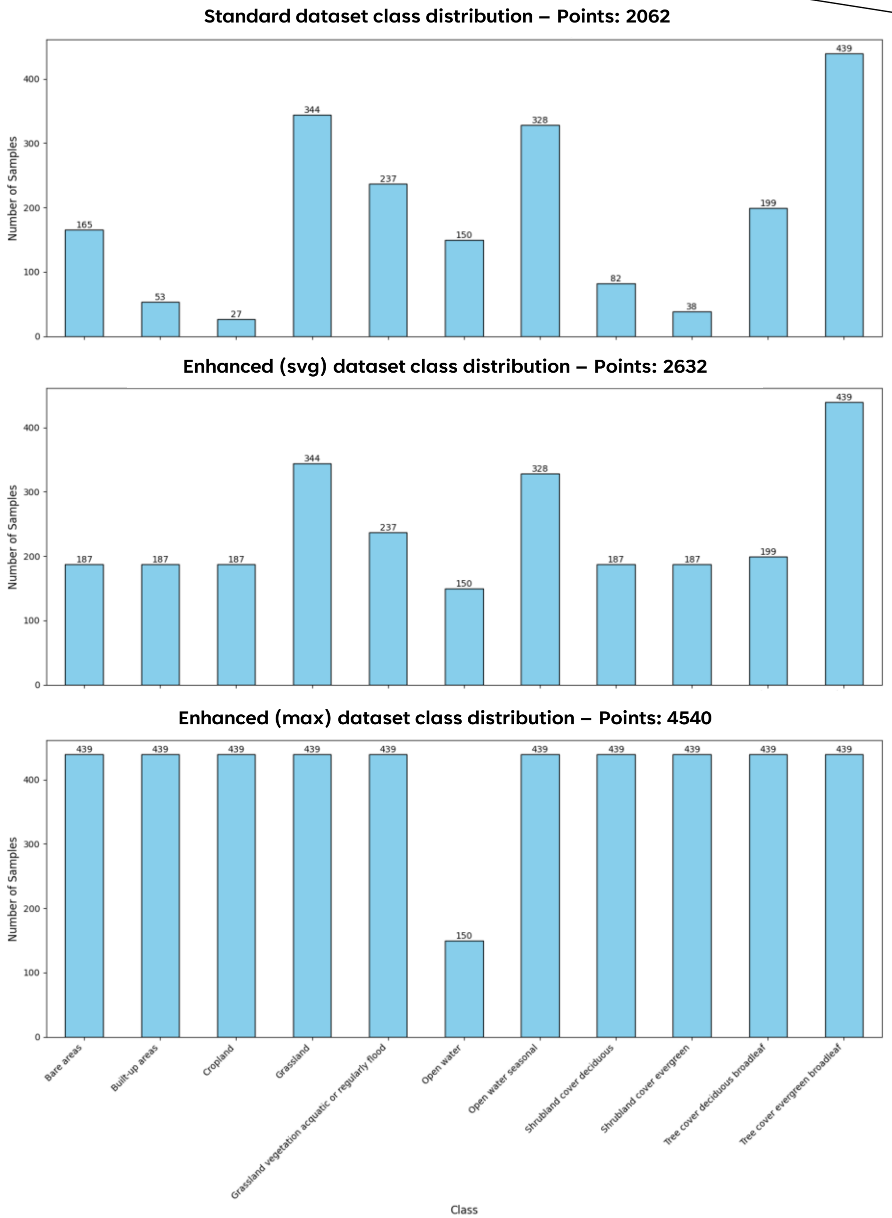
All features are computed from a single monthly composite and stored in a structured format for downstream dataset generation.

FOURTH COMPONENT: DATASET GENERATION

This component constructs the training and evaluation datasets by combining the extracted features with annotated land cover labels. It supports multiple data sources and configurable sampling strategies.

The main steps include:

- **Loading features** from Sentinel-2 composites, DEMs, and generated indices;
- **Extracting class labels** from ground truth shapefiles and/or existing land cover maps;
- **Mapping spatial coordinates to pixel indices**, ensuring geospatial alignment;
- **Sampling strategies** to balance class distributions, including:
  - *Standard*: using only annotated shapefiles;
  - *Enhanced*: combining shapefiles with additional points from existing LC maps;
  - *FullLC*: sampling entirely from LC maps, keeping shapefile points for evaluation;
- **Saving structured datasets** (CSV) with labeled samples and diagnostic reports;



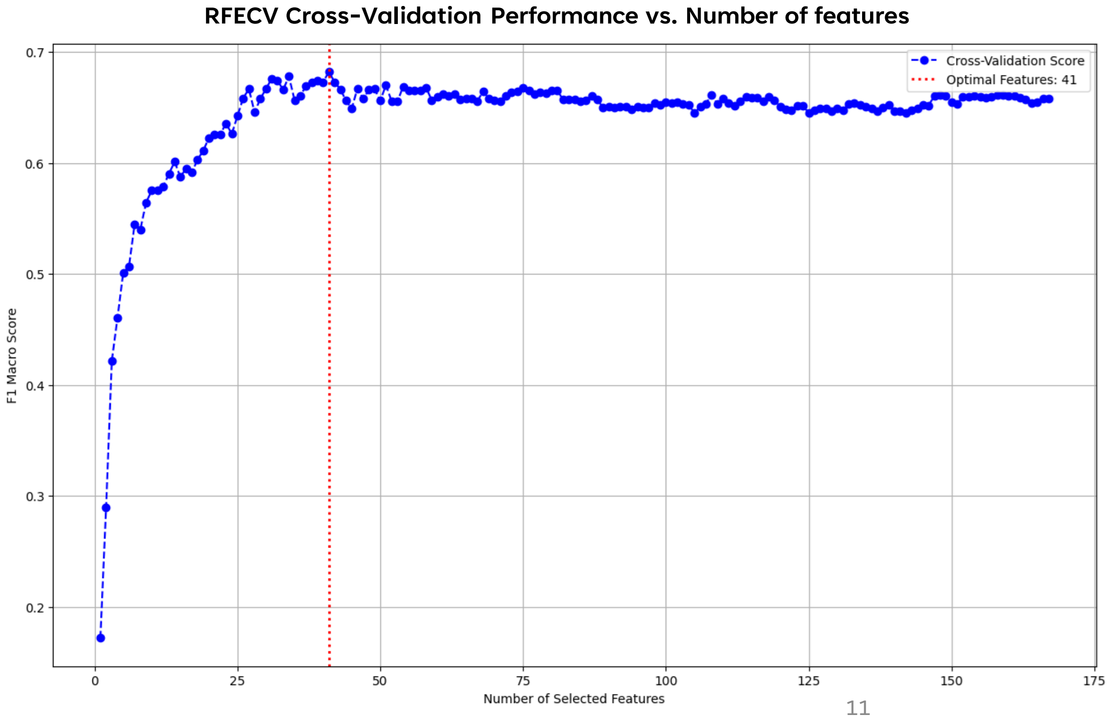
FIFTH COMPONENT – SVM PIPELINE

Features preprocessing

- ***Standardization***: centers data with zero mean and unit variance;
- ***Min-Max Scaling***: rescales data to a [0, 1] range
- **Quantile-based scaling** is also supported for robustness against outliers.

Feature Selection

1. ***Recursive Feature Elimination with Cross-Validation (RFECV):*** An iterative, supervised method that evaluates the impact of each feature on SVM model performance. At each iteration, the least important features are removed based on the model's coefficients, and cross-validation is used to identify the optimal subset that maximizes classification performance (macro F1-score)..

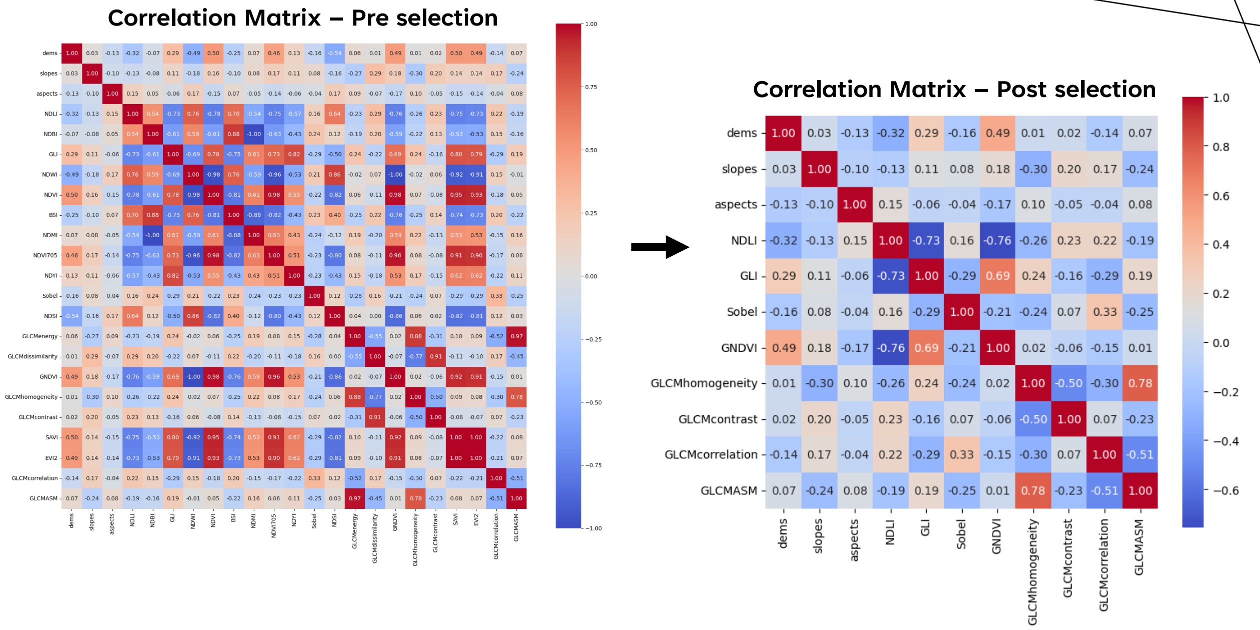


SVM PIPELINE – FEATURES SELECTION AND REDUCTION

Feature Selection

2. **Multicollinearity Analysis:** an unsupervised technique designed to remove redundancy among features. It identifies highly correlated feature pairs using a correlation matrix and selects which features to retain based on a hybrid ranking metric. This helps preserve informative features while reducing noise caused by multicollinearity.

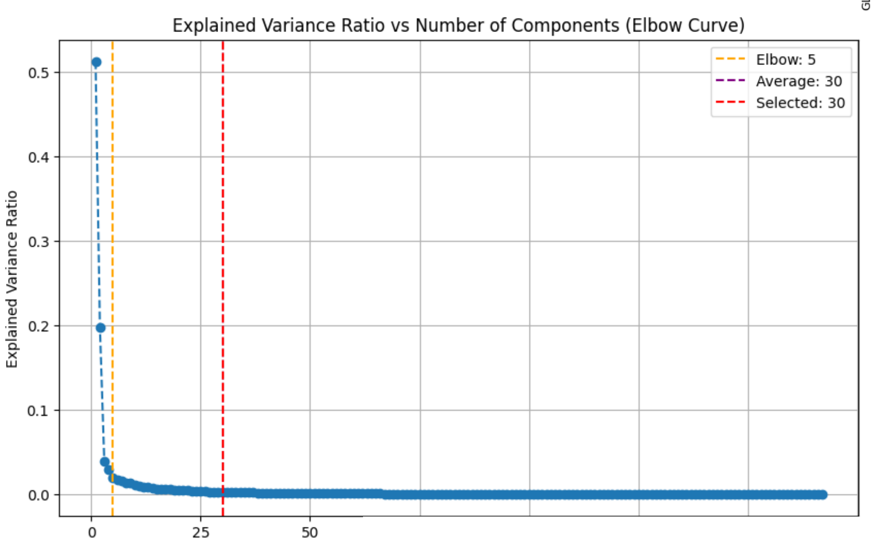
3. **No selection strategy** is also an option during this phase.



Feature Reduction

Dimensionality can be further reduced using **Principal Component Analysis (PCA).** The selected components retain most of the original variance and improve efficiency. Three strategies are available for choosing the number of components:

- Cumulative explained variance threshold;
- Elbow method;
- Average of both criteria



SVM PIPELINE - TRAINING APPROACH

This component supports two parallel strategies for training SVMs.

Binary SVMs (One-vs-All)

- One binary classifier is trained and calibrated per class
- Hyperparameters are tuned independently for each model
- Calibrated outputs can be aggregated using softmax-based normalization or a standard approach scikit-learn like.

Multiclass SVM (One-vs-Rest)

- A single model is trained using scikit-learn's SVC in a One-vs-Rest configuration
- Hyperparameters (C, γ) are selected via grid search with 3-fold cross-validation
- After training, probability calibration is applied using isotonic regression

Evaluation

- Models are evaluated using accuracy, macro F1-score, and class-wise precision/recall
- Calibration curves are used to assess probability reliability

	Accuracy	Macro F1-Score	Macro Recall	Macro precision
OVO SVM not calibrated	80%	77%	77%	78%
OVO SVM calibrated	80%	71%	70%	74%
Binary SVMs	81%	69%	67%	76%

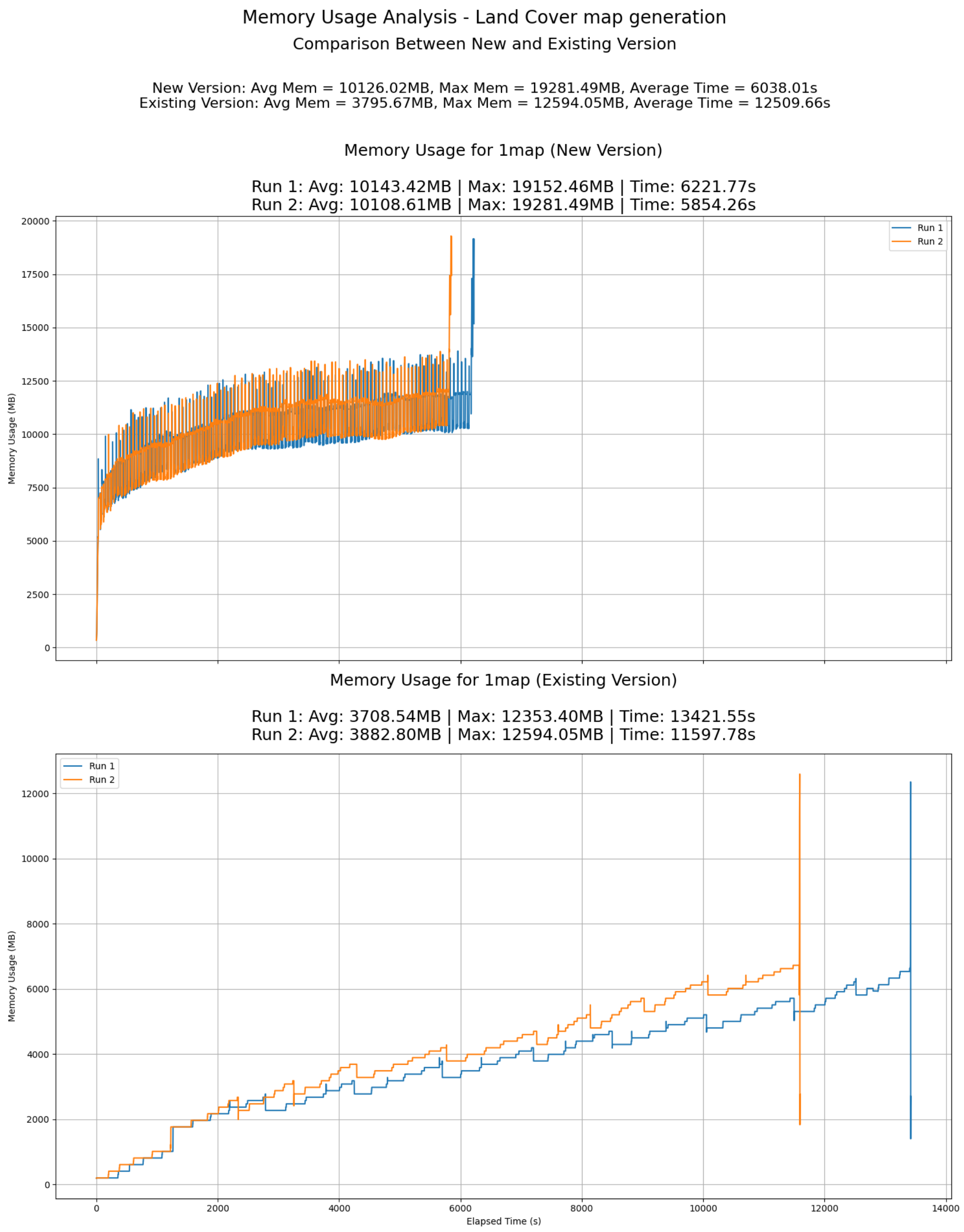
SIXTH COMPONENT: LAND COVER MAP GENERATION

The final step applies the trained SVMs to classify each pixel in the input tile using a patch-based approach.

- **Input**: Same features used to train the employed SVMs;
- **Process**ed in spatial chunks using xarray and dask;
- **Outputs**:
  - GeoTIFF maps with predicted classes and class probabilities;
  - PNG export for visual inspection;

Performance evaluation

1. The reimplemented pipeline significantly reduces execution time, cutting processing duration by nearly 50%, while increasing memory usage due to parallelization and full-scene loading with xarray.
2. Despite the higher memory footprint, the new approach is more efficient and scalable, especially when generating multiple maps in a single run.

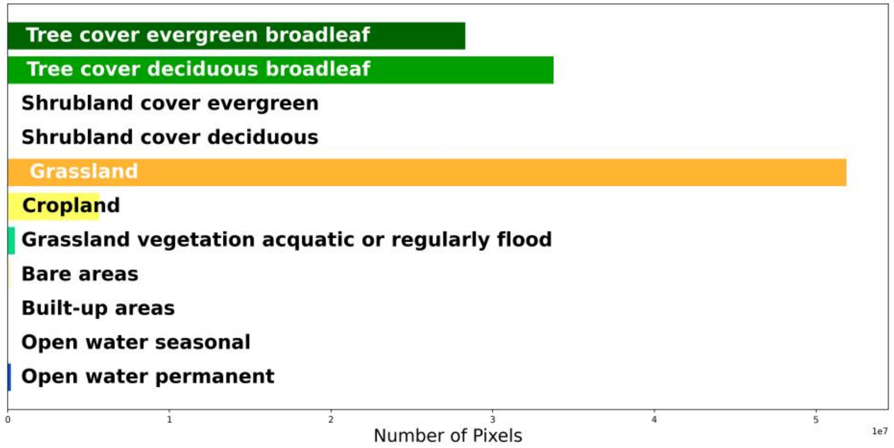


SVM PIPELINE - TRAINING APPROACH

- **Tile 21KUQ, 2019.**
- **Most predicted class**: *Grassland* **EP** and *Tree cover evergreen broadleaf* in **RP**.
- **Prediction distribution**: concentrated on few classes in **EP** and more balanced in the **RP**.
- **Minor classes**: more detected in **RP**

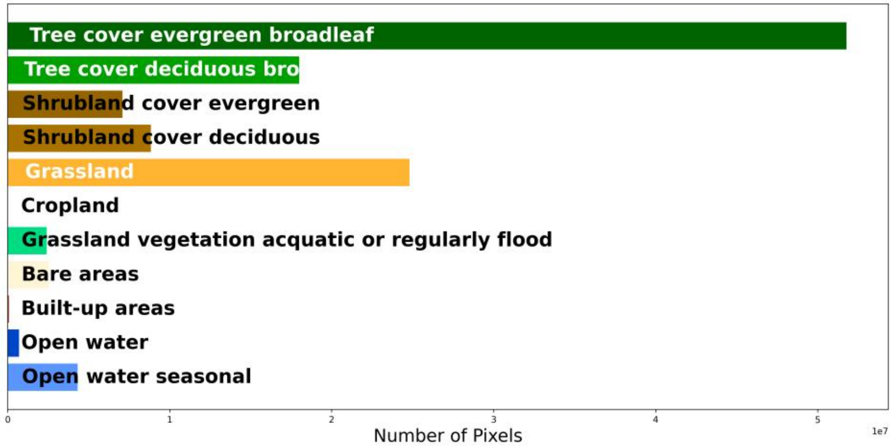
21KUQ 2019 – LC map with existing pipeline (**EP**)

Predicted classes distribution



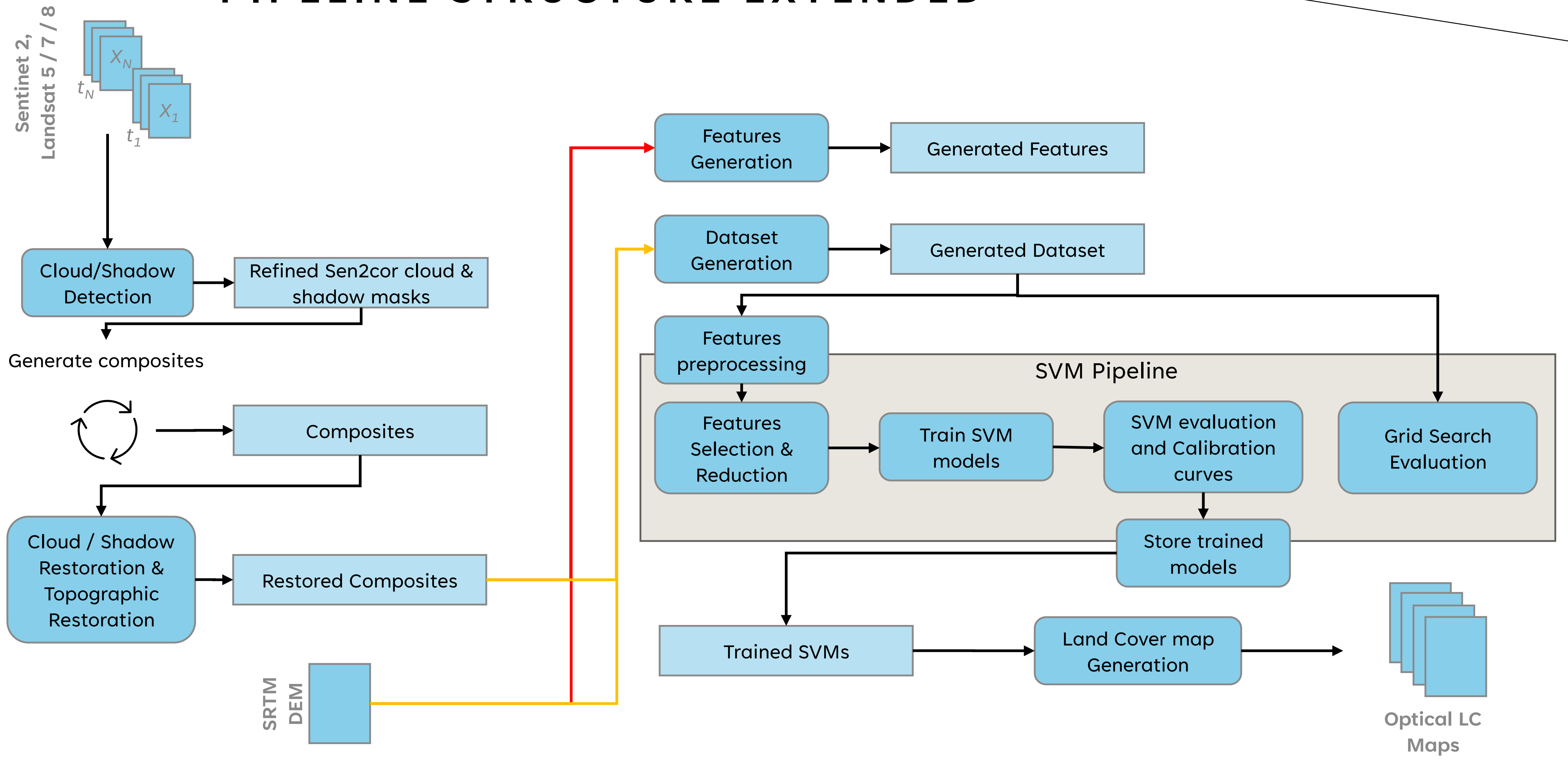
21KUQ 2019 – LC map with reimplemented pipeline (**RP**)

Predicted classes distribution



THANK YOU

PIPELINE STRUCTURE EXTENDED



USED FEATURES

Group	Features
Vegetation indeces	Normalize Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), Red-edge Normalized Difference Vegetation Index (NDVI705), Green Leaf Index (GLI), Enhanced Vegetation Index 2 (EVI2), Soil-Adjusted Vegetation Index (SAVI), Normalized Difference Yellow Index (NDYI), Normalized Difference Lignin Index (NDLI)
Moisture and water content	Normalized Difference Moisture Index (NDMI), Normalized Difference Water Index (NDWI), Normalized Difference Snow Index (NDSI)
Soil and built-up surface indicators	Normalized Difference Built-up Index (NDBI), Bare Soil Index (BSI)
Structural and edge features	Sobel Edge, GLCM features, DEM, Slope and aspect
Spectral	Bands from composites

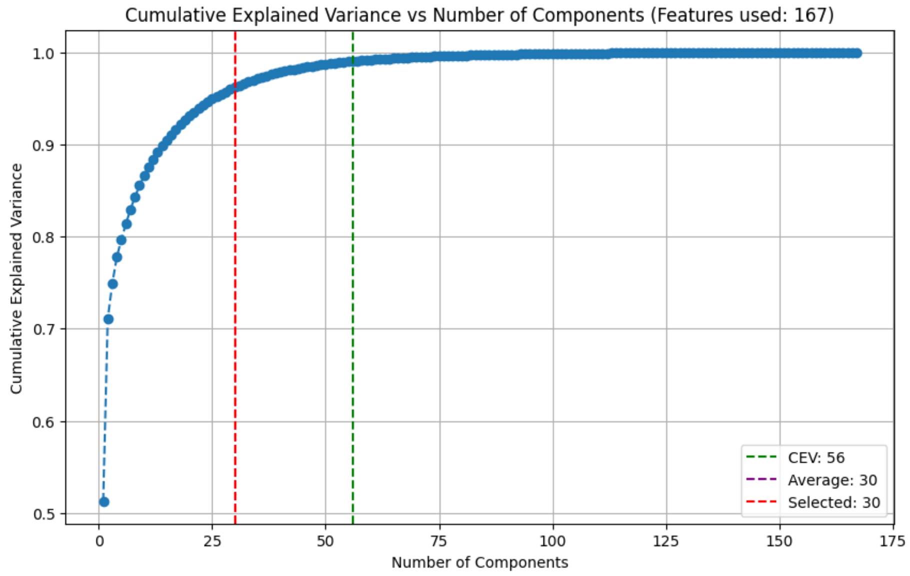
SVM PIPELINE – FEATURES REDUCTION

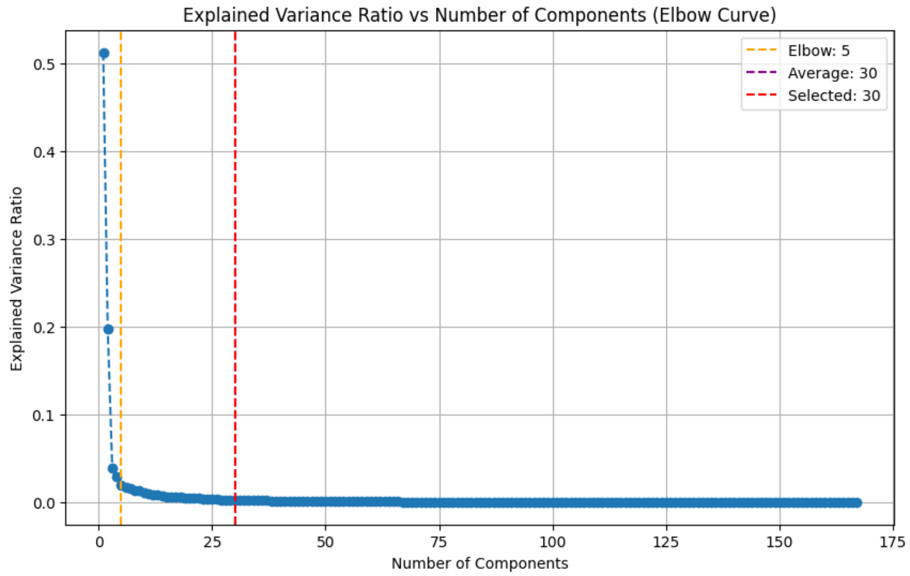
Dimensionality is further reduced using **Principal Component Analysis (PCA)**.

The selected components retain most of the original variance and improve efficiency.

Three strategies are available for choosing the number of components:

- Cumulative explained variance threshold;
- Elbow method;
- Average of both criteria.





SVM PIPELINE – MULTICOLLINEARITY ANALYSIS