# Review document on Exploring the use of Artificial Intelligence (AI) to Optimize the Exploitation of satellite EO and modelling data

## Project Identification

| | |
|---|---|
| **Project Full Title** | Water scenarios for Copernicus Exploitation |
| **Project Acronym** | Water-ForCE |
| **Grant Agreement** | 101004186 |
| **Starting date** | 01.01.2021 |
| **Duration** | 36 months |

## Document Identification

| | |
|---|---|
| **Deliverable number** | D5.3 |
| **Deliverable Title** | Exploring the use of Artificial Intelligence (AI) to Optimize the Exploitation of satellite EO and modelling data. |
| **Type of Deliverable** | Report |
| **Dissemination Level** | Public (PU) |
| **Work Package** | WP5 |
| **Leading Partner** | IHE (WP5) – Antea Group (WP5.3) |

## History of Changes

| Date | Version | Comments |
|---|---|---|
| 18/11/2022 | V1.0 | Thant, S., Henkens, S., Messens, F. |
| 12/12/2022 | V2.0 | Thant, S., Henkens, S., Messens, F. |
| | | |

Water-ForCE is a CSA that has received funding form the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101004186.

## List of Acronyms

| | |
|---|---|
| **AB** | Advisory Board |
| **AGA** | Annotated Grant Agreement |
| **CA** | Consortium Agreement |
| **CSA** | Coordination and Support Action |
| **CT** | Coordination Team |
| **DoA** | Description of Action |
| **DMP** | Data Management Plan |
| **EB** | Executive Board |
| **EC** | European Commission |
| **IPR** | Intellectual Property Rights |
| **FA** | Funding Authority |
| **GA** | Grant Agreement |
| **GAs** | General Assembly |
| **GPDR** | General Data Protection Regulation |
| **PIP** | Project Implementation Plan |
| **PM** | Person-months |
| **PO** | Project Officer |
| **SDGs** | Sustainable Development Goals |
| **TL** | Task Leader |
| **WG** | Working Group |
| **WP** | Work Package |

# Abstract

Task 5.3 of the Copernicus project aims at providing a state-of-the-art literature overview on the use of artificial intelligence and machine learning in processing EO data. It provides information on the ways Artificial Intelligence (AI) can support end user needs and recommendations on the use of AI to optimize the exploitation of satellite EO and modelling data in general are given. The number of projects and initiatives focusing on the need of the integration of EO data processing, -assimilation and application building and artificial intelligence techniques is significant and rising. Important bottlenecks limiting the use of AI for optimal exploitation of EO data are the lack of labeled datasets, the volume of data and the explainability/causality of events. Among others, streamlined platforms and initiatives and a focus on a holistic approach for the implementation of AI for EO are recommendations for guiding different stakeholders to the for them relevant information.

# Table of Contents

# 1 Introduction

## 1.1 Water-ForCE

The **Horizon2020** project Water-ForCE (Water scenarios For Copernicus Exploitation) will develop a Roadmap for Copernicus **Inland Water Services**, aiming to better integrate the entire inland **water cycle within the** Copernicus Services. It will address current disconnects between remote sensing / in-situ observation and the user community. Clarity in terms of the needs and expectations of both public and private sectors, as well as the wider research and business innovation opportunities will be delivered. The Roadmap will advise on a strategy to ensure effective uptake of water-related services by end users and further support the implementation of relevant directives and policies.

The Water-ForCE consortium is led by the University of Tartu (Estonia) and consists of 20 organisations from all over Europe. It connects experts in water quality and quantity, in policy, research, engineering and service sectors. Through close collaborations with these communities, Water-ForCE will:

- Analyse EU policies to identify where the Copernicus Services can improve monitoring programs and how the Copernicus data can be more effectively used in developing and delivering the next versions of the directives.
- Specify the requirements for future Copernicus missions (e.g. optical configuration of Sentinel-2E and onward, hyperspectral sensors).
- Optimize future exploitation for inland water monitoring & research and, consequently, (a) enlarge the service portfolio and (b) improve the performance of current Services.

The project is divided in eight work packages (WP), each of them focusing on a specific problem and/or target of the Copernicus Service (Figure 1). The following report is part of **WP5** which focusses on **Modelling and Data assimilation**.

## 1.2 Context WP5

WP5 aims to augment the knowledge acquired in WP1-WP4 by identifying the potential for future use of different satellite EO techniques in modelling of water resources for support of decision makers towards adaptive management of water resources and policy implementation.

Current issues in Copernicus hindcast/forecast capabilities will be assessed and recommendations for future services will be provided. Furthermore, the value of satellite EO data to modelling will be examined and the use of Artificial Intelligence (AI) to optimize the exploitation of satellite EO and modelling data considered. Finally, WP5 will provide insights on how the integration of satellite EO and coinciding modelling aspects can lead to beneficial policy support and decision making.

Detailed work package tasks/milestones include:

- Report on needs assessment for Copernicus EO needs for modelers and decision makers, including an overview of the main stakeholders identifying these needs.
- Technical recommendations report on Copernicus services and the related data in order to improve the monitoring and modelling of water bodies.
- State of the art and recommendations on the use of Artificial Intelligence (AI) to optimize the exploitation of satellite EO and modelling data.
- Report on integration of satellite EO and modelling aspects for providing better decision support and operational management, including recommendations.

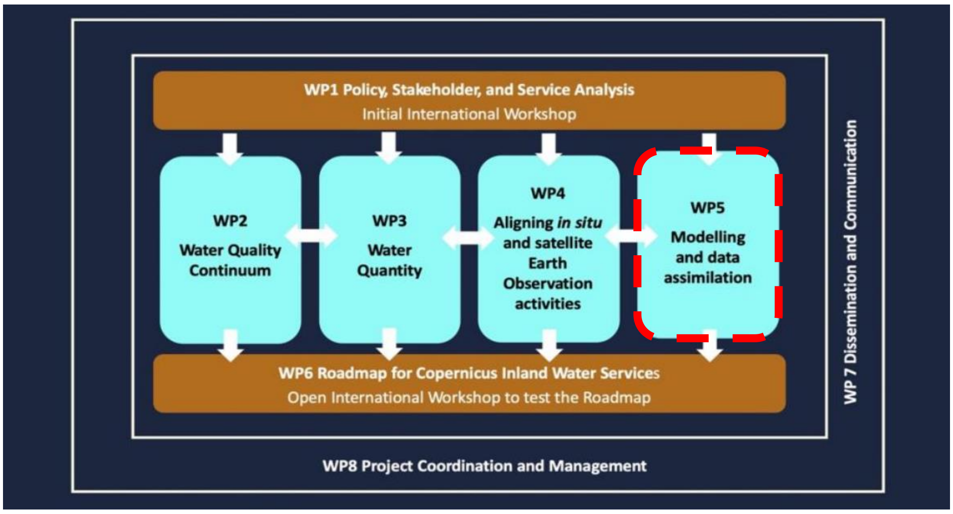

WP1 Policy, Stakeholder, and Service Analysis
Initial International Workshop
WP2
Water Quality Continuum
WP3
Water Quantity
WP4
Aligning *in situ* and satellite Earth Observation activities
WP5
Modelling and data assimilation
WP6 Roadmap for Copernicus Inland Water Services
Open International Workshop to test the Roadmap
WP 7 Dissemination and Communication
WP8 Project Coordination and Management

*Figure 1: Schematic overview on the different work packages in the Water-ForCE project, concluding in a Roadmap for Copernicus Inland Water Services.*

## 1.3 Objectives & approach T5.3

Task 5.3 of the Copernicus project aims at providing a state-of-the-art literature overview on the use of artificial intelligence and machine learning in satellite EO data assimilation and -modelling. Information on the ways of how AI can support end user needs will be provided. However, the implementation of AI for EO practices is on the rise, generating numerous amounts of papers and review articles. Tackling all user needs (identified during previous Water-ForCE deliverables) or opportunities in which AI can play a role and discussing which AI algorithms should be used would therefore be impractical and out of scope. Good and detailed review papers already exist on how AI can be applied for the majority of these topics.

Therefore, ***this deliverable will handle the topic of AI for EO in general and a more conceptual framework***, focusing on providing *an overview on the current status of the use of AI by the EO community (their attitude towards implementing AI) and ideas/visions on how AI can be of further support to the needs of the end user.* The latter *indicating bottlenecks* for the end user, limiting the implementation of AI for EO purposes and therefore *limiting the optimal exploitation of EO data*. Specific questions coming to mind are:

- What is the state of the art on applying AI in EO domains? In which EO domains is AI used?
- How can AI support in solving specific end user needs such as accurately correcting systematic forecast errors and predict the time evolution of geophysical parameters from satellite and other geophysical inputs? Different approaches can be given.
- Which AI related prospects and needs for developing and deploying next–generation observation, data assimilation, data processing, and modelling for environmental applications can be identified?
- What are the current bottlenecks limiting the use of AI for EO?

In this report, chapter 2 will explain the different definitions used when entering the world of artificial intelligence. An overview on the implementation of AI in various EO domains or projects and how it supports achieving today's UN Sustainable Development Goals (SDGs) is provided in chapter 3. An overview on AI/ML approaches currently used in the (pre-) processing of EO data (including challenges & pitfalls) is shown in section 4. In section 5 different types of end user needs of the EO community and the usage of AI in order support the exploitation of EO data are discussed. Along different AI strategies and techniques also bottlenecks with regard to AI implementation and the Technological Readiness Level of techniques and applications is considered.

Conclusions and recommendations can be found in section 6.

# 2 Introduction of AI, ML and DL

Artificial intelligence, machine learning and deep learning are distinctive terms, related to one another as shown in Figure 2. In general, AI is described as the capability of a machine to act as a human-being, mimicking human intelligence, by performing tasks such as object detection & recognition and displaying skills in problem solving, learning, planning.Machine learning is considered a subsection of AI, based on the idea that ML models can learn from input data ("training") by which they can identify patterns and make predictions and/or decisions. Deep learning is a large subdomain of machine learning considering highly complex algorithms for an increased degree of abstraction. It can handle a wide variety of input data and learning architectures and is based on the machine learning concept of neural networks.

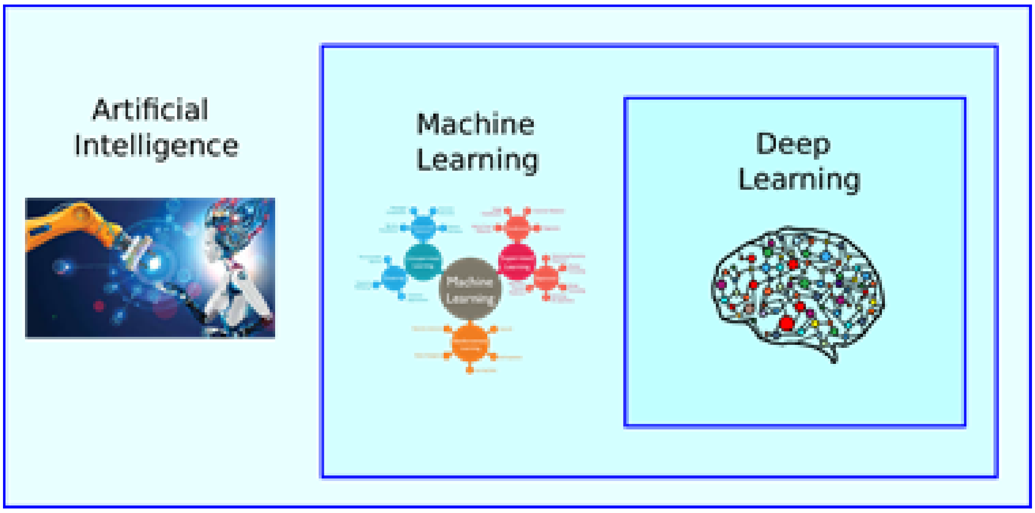


*Figure 2: The relation between artificial intelligence, machine learning and deep learning.*

The difference between ML and classical data programming is that machine learning is a paradigm for creating models based on example data. The relation between input and desired output data is "learned" (trained) instead of directly programmed (Figure 3).

The relation between the input and output is referred to as the 'function' or in machine learning often as the 'model'. Hence machine learning should not only be considered a new tool in the toolbox of a researcher, but rather as another paradigm on how to solve problems and understand information. This method of empirical model building typically has an advantage over classical 'forward' model building when large amounts of data are available and the relations between input and output data are too complex to directly identify and/or program.

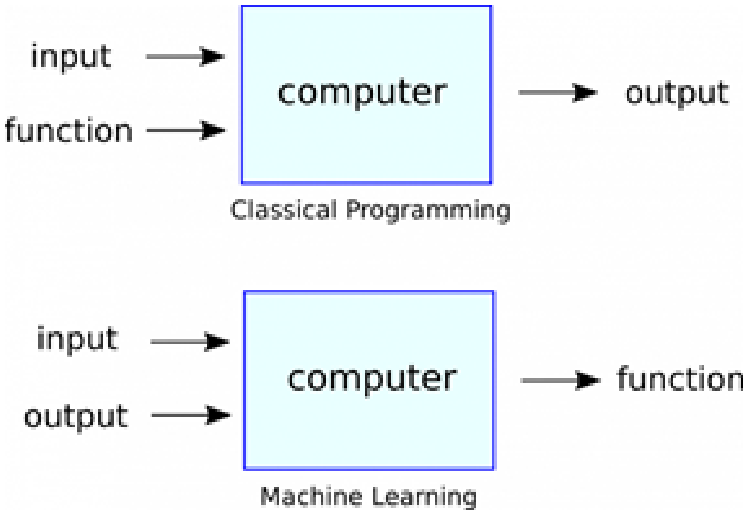


*Figure 3: Machine learning vs. classical modeling and the programming paradigm.*

Machine learning consists of several learning methods:

- ***Supervised learning***: in supervised learning the model is provided with labeled training datasets. The algorithm learns the relationship between input-output variables, and can learn over time, improving its accuracy. The requirement of labeled datasets depicts the need of human interaction (Ennouri *et al.*, 2021; Reichstein *et al.*, 2019).
- ***Unsupervised learning***: unsupervised methods rely solely on non-labeled training data, it is up to the algorithm to depict underlying structures and deduce patterns and associations amongst the data. The goal is to get insights in the dataset, without predefining the nature of these relationships (Ennouri *et al.*, 2021; Reichstein *et al.*, 2019).

- ***Semi-supervised learning:*** method using a large amount of unlabeled data, combined with a small amount of labeled data (Reichstein *et al.*, 2019), increasing the learning accuracy (weak supervision), while limiting the high costs of extensive labeled datasets.
- ***Reinforced learning***: method using trial and error technique in order to learn making the correct decisions (Ennouri *et al.*, 2021), based on the use of reward rules.

# 3 AI in the world of Earth Observation

Artificial Intelligence (AI), Cybersecurity, Internet of Things, Big Data, High Performance Computing, 5G, and Software are the main digital specializations for Europe's digital transformation by 2030 (Europe's Digital Decade, 2021) (Water-ForCE, 2022b).

The Euroconsult report (2021) identifies AI, machine learning (ML) or cloud-computing as trending technologies and enablers for EO applications and services. AI to EO computer vision applications is one of the biggest contributors, with learning algorithms able to reduce the error rate of detection or identification. An important aspect of ML is the availability of sufficient training datasets which increases the quality of the algorithms and results overall. For instance, for water quality assessment, the ML needs sufficient in-situ data which is often not available. Projects like MONOCLE or Hypernets are trying to fill the gap of open ground truth/training datasets. Other AI applications to EO include data processing, change detection, object recognition, identification, prediction and so on (Water-ForCE, 2022b).

The rising amount of governmental and private projects and initiatives indicate the significant value of AI towards EO applications and services.

An overview (non-limitative list) of the most important AI for EO related projects is presented in §3.1, providing also some specific applications making use of AI techniques. Section 3.2 discusses the impact of using AI for EO in achieving the UN Sustainable Development Goals (SDG).

## 3.1 Projects and applications implementing AI for EO

Today, a high variety of projects concerning the use of artificial intelligence for the exploitation of Earth Observation exist. Here, a (non-limitative) list of important projects with regard to the implementation of AI for societal challenges, related to EO, is given:

- AI4Copernicus: a very important H2020 project, bridging the worlds of Artificial Intelligence (AI) and Earth Observation (EO) by reinforcing the AI4EU AI-on-demand platform with datasets, tools and services relevant to Copernicus data. This in order to enhance the uptake of EO resources by users of various socio-economic domains, e.g. Agricultural sector, Health sector, Security sector, ... AI4EU resources will be made accessible on EO data platforms (DIAS).

- AI4EU: the AI on-demand platform of the European Union provides access to expert AI knowledge (incl. research), technologies, applications & tools and experts in the domain. They provide AI assets (algorithms, datasets, models,...) which can be tested on the AI4EU Experiments platform, a platform which can be used to build your own AI based solutions.
- GEO.INFORMED: a 4-year project funded by the Flemish Reasearch Foundation (FWO) aiming to develop deep learning workflows able to transform Sentinel 2 satellite data into ready-to-use data products for environmental policy agencies.
- Earth Science Data Systems Program (NASA): programme focusing on the use of AI to increase the capability of data systems, enhance the performance of operations and maximize the exploitation of the NASA Earth observing data.
- EO4society (ESA): "*It pioneers new EO services and scientific discoveries, stimulating downstream industry growth, and supporting international responses to global societal challenges*". The program offers workshops/training in the use of AI for EO services.
- Planetary computer (Microsoft): the planetary computer provides a catalog on global environmental data and a variety of applications for analysis and accessing actionable information.
- H2020 projects: a lot of projects incorporating the use of AI for earth and environmental sciences have been launched, a small selection:
  - ARTIST ARTificial Intelligence for Seasonal forecast of Temperature extremes
  - CLINT CLImate INTelligence: Extreme events detection, attribution and adaptation design using machine learning
  - Xaida extreme events: artificial intelligence for detection and attribution
  - AIDA Artificial Intelligence Data Analysis
  - ...

Due to the amount of H2020 projects they are not all described in detail, only the, into our opinion, most relevant ones. More information on e.g. the projects above can be obtained by clicking the link.

- CENTURION: a EU project developing ground-breaking advances in Big Data and Artificial Intelligence, creating tools to give consistent, straightforward access to EO Analysis-Ready Data and AI analytics, for use by experts and non-experts alike.
- DeepCube: DeepCube is a Horizon 2020 Space project unlocking the potential of big Copernicus data with Artificial Intelligence and Semantic Web technologies, with the objective to address problems of high environmental and societal impact.
- CALLISTO: a H2020 project, aiming to bridge the gap between Copernicus Data and Information Access Services (DIAS) providers and application end users through dedicated Artificial Intelligence (AI) solutions. It will provide an interoperable Big Data platform integrating Earth Observation (EO) data with crowdsourced and geo-referenced data.
- AI4SAR: ICEYE Analytics builds AI/ML applications for heavy-duty image processing and scalable analytics. ML based applications are built for SAR processing, time series analysis and data handling.

As numerous projects are focusing on the implementation of AI/ML in EO data handling and -processing also the amount of ML based applications is rising. Some specific applications are highlighted:

- 52north: application built for the detection of permanent water surfaces based on SAR data to help increase the accuracy of flood detection. Solving the issue of a lack of optical data that often occurrsduring heavy rainfall events.
- ESA WorldCover: Land cover and land use classification products are among the most important when addressing environmental problems. Often spatial resolution is limited for specific use cases and significant amount of time passes between product updates. ESA now provides a 10 m resolution global land cover map, which is based on ML algorithms and reduces the data processing to less than 5 days.

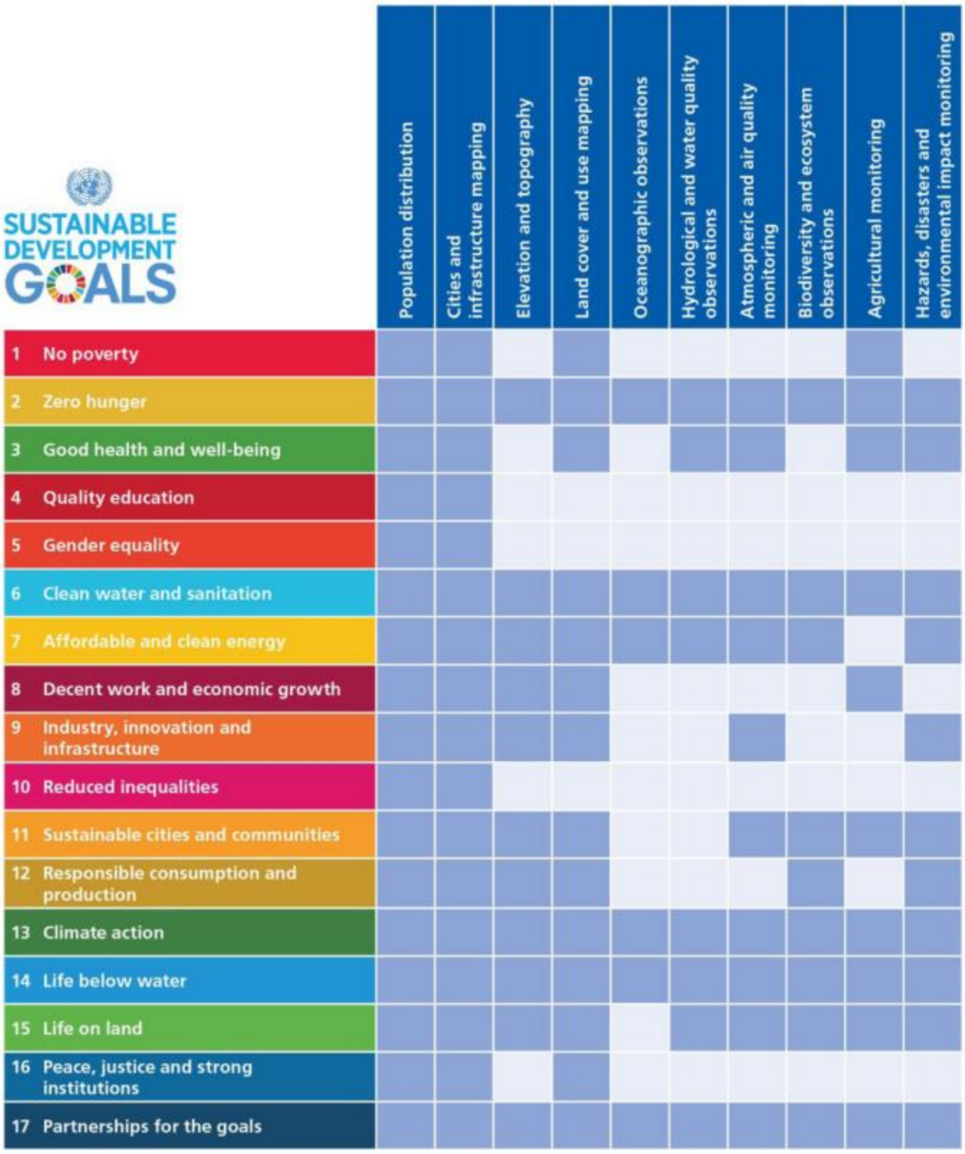
- Destination Earth: At the heart of the DestinE project is the concept of the digital twin – a virtual Earth that simulates natural processes and human activity. Observing such replicas will help researchers understand change and help shape policies to mitigate extreme climate-related risks to society. Artificial intelligence (AI) and machine learning will make this interactive framework more flexible, efficient and faster.

## 3.2 AI to support achieving SDG's

As stated earlier, the implementation of AI in various domains of EO and remote sensing has proven to be of significant value and has been rapidly advancing the past decades. Machine learning techniques have become a vast component in the processing of remote sensing data, supporting Earth Observation approaches. EO and RS data is most often characterized by high spatial- and temporal resolutions and coverage, making them a viable tool for the indicators monitoring and measuring the progress towards the UN 17 Sustainable Development Goals (GEO, 2017).

Figure 4 provides an overview on the type of EO data by which the 17 SDG's can become measurable in a more quantitative way. More information on how EO data can contribute in monitoring the SDG's, complemented with specific case studies, can be found in GEO (2017). Whereas Andries *et al.* (2018) illustrates how EO data can be translated into sustainable development indicators. It became clear also from D1.6 of the Water-ForCE project that Copernicus Services hold an important role in helping to achieve the UN's SDGs. The analysis in Water-ForCE D1.6 identified the possible links between EO parameters specific for the inland water quality and quantity and 12 of the UN Sustainable Development Goals (Water-ForCE, 2022). Pahlevan *et al.* (2022) discuss how EO can be used for monitoring inland water quality and consistent reporting of SDG 6.3.2/6.6.1 indicators.

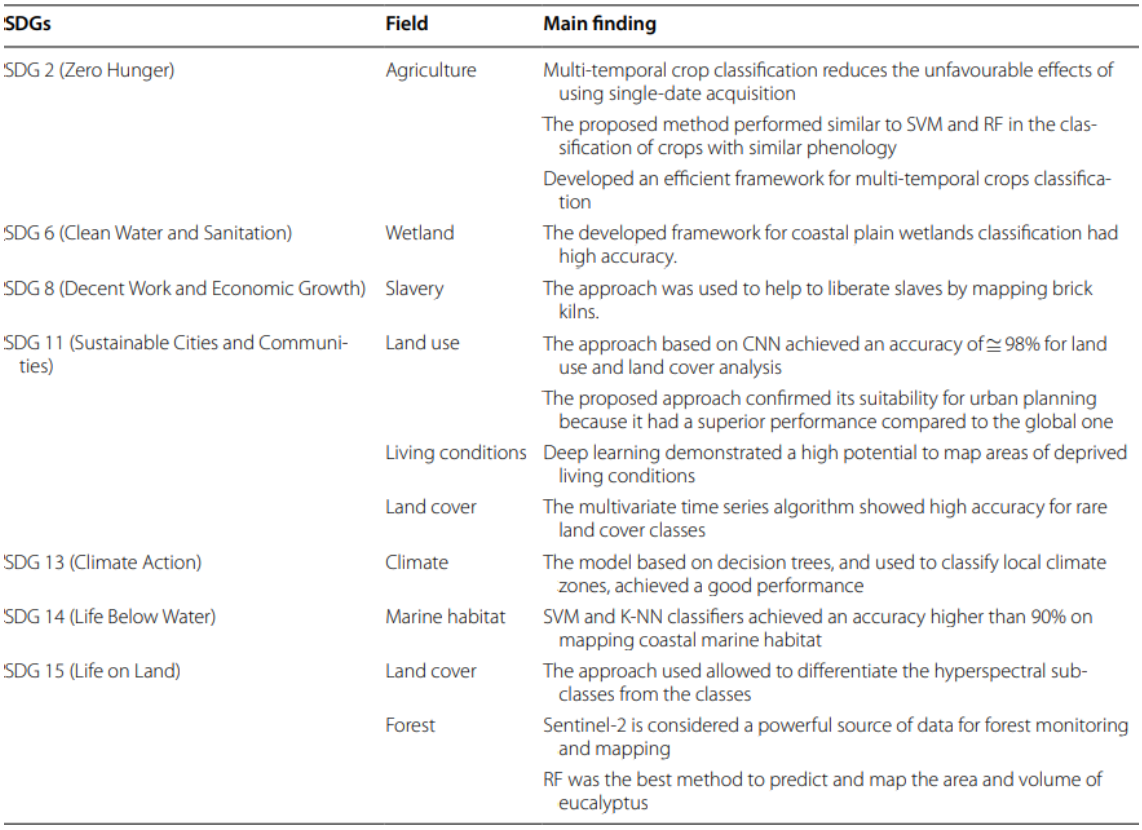
Consequently, ML is believed to also have a considerable potential in supporting the achievement of the different SDG's (Ferreira *et al.*, 2020). Ferreira *et al.* (2020) discusses different ML algorithms, being among the most relevant ML techniques currently used in the support of EO data analysis and processing (see also §4.1). Furthermore, Ferreira *et al.* (2020) also provides overviews of applications of classification, clustering, regression and dimension reduction techniques, making use of EO data, with regard to the different SDG’s (Table 1).



| SUSTAINABLE DEVELOPMENT GOALS | Population distribution | Cities and infrastructure mapping | Elevation and topography | Land cover and use mapping | Oceanographic observations | Hydrological and water quality observations | Atmospheric and air quality monitoring | Biodiversity and ecosystem observations | Agricultural monitoring | Hazards, disasters and environmental impact monitoring |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 No poverty | ● | ● | | ● | | | | | ● | |
| 2 Zero hunger | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 3 Good health and well-being | ● | ● | | ● | | ● | ● | | ● | ● |
| 4 Quality education | ● | ● | | | | | | | | |
| 5 Gender equality | ● | ● | | | | | | | | |
| 6 Clean water and sanitation | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 7 Affordable and clean energy | ● | ● | ● | ● | ● | ● | ● | ● | | ● |
| 8 Decent work and economic growth | ● | ● | ● | ● | | | | | ● | |
| 9 Industry, innovation and infrastructure | ● | ● | ● | ● | | | ● | | | ● |
| 10 Reduced inequalities | ● | ● | | | | | | | | |
| 11 Sustainable cities and communities | ● | ● | ● | ● | | | ● | ● | ● | ● |
| 12 Responsible consumption and production | ● | ● | ● | ● | | | | ● | | ● |
| 13 Climate action | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 14 Life below water | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 15 Life on land | ● | ● | ● | ● | | ● | ● | ● | ● | ● |
| 16 Peace, justice and strong institutions | ● | ● | | ● | | | | | | |
| 17 Partnerships for the goals | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |

*Figure 4 : Overview on types of  EO data (columns) applicable for monitoring each of the 17 Sustainable Development Goals (Source: GEO, 2017).*

*Table 1: Examples of application of classification methods towards SDGs using EO data (Source: Ferreira et al., 2020)*



| SDGs | Field | Main finding |
|---|---|---|
| SDG 2 (Zero Hunger) | Agriculture | Multi-temporal crop classification reduces the unfavourable effects of using single-date acquisition |
| | | The proposed method performed similar to SVM and RF in the classification of crops with similar phenology |
| | | Developed an efficient framework for multi-temporal crops classification |
| SDG 6 (Clean Water and Sanitation) | Wetland | The developed framework for coastal plain wetlands classification had high accuracy. |
| SDG 8 (Decent Work and Economic Growth) | Slavery | The approach was used to help to liberate slaves by mapping brick kilns. |
| SDG 11 (Sustainable Cities and Communities) | Land use | The approach based on CNN achieved an accuracy of ≅ 98% for land use and land cover analysis |
| | | The proposed approach confirmed its suitability for urban planning because it had a superior performance compared to the global one |
| | Living conditions | Deep learning demonstrated a high potential to map areas of deprived living conditions |
| | Land cover | The multivariate time series algorithm showed high accuracy for rare land cover classes |
| SDG 13 (Climate Action) | Climate | The model based on decision trees, and used to classify local climate zones, achieved a good performance |
| SDG 14 (Life Below Water) | Marine habitat | SVM and K-NN classifiers achieved an accuracy higher than 90% on mapping coastal marine habitat |
| SDG 15 (Life on Land) | Land cover | The approach used allowed to differentiate the hyperspectral subclasses from the classes |
| | Forest | Sentinel-2 is considered a powerful source of data for forest monitoring and mapping |
| | | RF was the best method to predict and map the area and volume of eucalyptus |

# 4 State of the art – literature review

## 4.1 Different AI/ML approaches in EO

Over the years, the use of ML algorithms and applications in different fields of EO and remote sensing has proven to be of significant value in terms of data-analysis and -assimilation. Machine learning techniques have been applied in various domains of EO, Table 2 giving a non-limitative list of examples. Sun & Scanlon (2019) indicate the importance of AI in environmental and earth and planetary sciences compared to other scientific domains (Figure 5).

Machine learning holds a lot of different techniques, each of them suitable for specific tasks in the EO realm. As explained in §2 ML in general makes use of supervised and unsupervised experiments. Ferreira *et al.* (2020) gives an overview on the most relevant algorithms currently being applied in RS, suitable for clustering and dimension reduction techniques in case of unsupervised experiments and classification and regression approaches when applying supervised operations.

The techniques mentioned above can be briefly described as followed (Holloway & Mengersen, 2018):

- ***Clustering***: technique to detect similarities between objects based on input variables and then classify the objects possessing similar characteristics into the same group (cluster) (unsupervised).
- ***Dimension reduction***: technique that reduces the number of input variables and ends with a set of variables capturing the most important information with regard to the original dataset (unsupervised).
- ***Classification***: technique used to allocate objects/phenomena to predefined groups and/or classes. The designation of an object to a group or class is based on the input variables that were given (supervised).
- ***Regression***: technique used to predict or estimate a response variable as a function of predictors (supervised).

A general overview on machine learning methods and algorithms can be found in Figure 6, whereas Table 3 gives an overview and brief explanation on relevant ML algorithms presently used in RS (EO) based on Ferreira *et al.* (2020), Holloway & Mengersen (2018) and TowardsDataScience (2022).

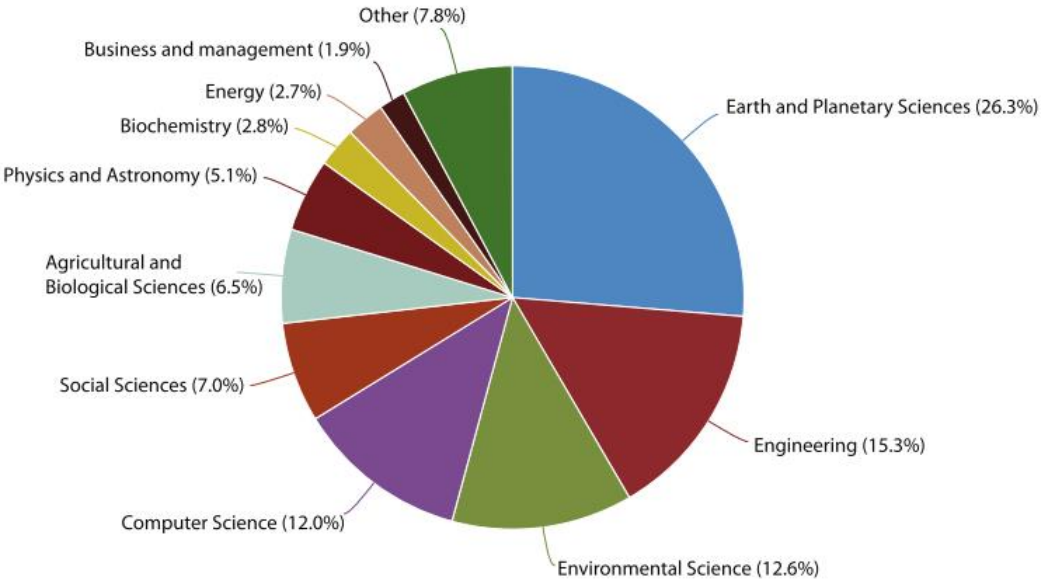


*Figure 5: Percentage of documents mentioning ML in the study of Sun & Scanlon (2019) by scientific domain.*

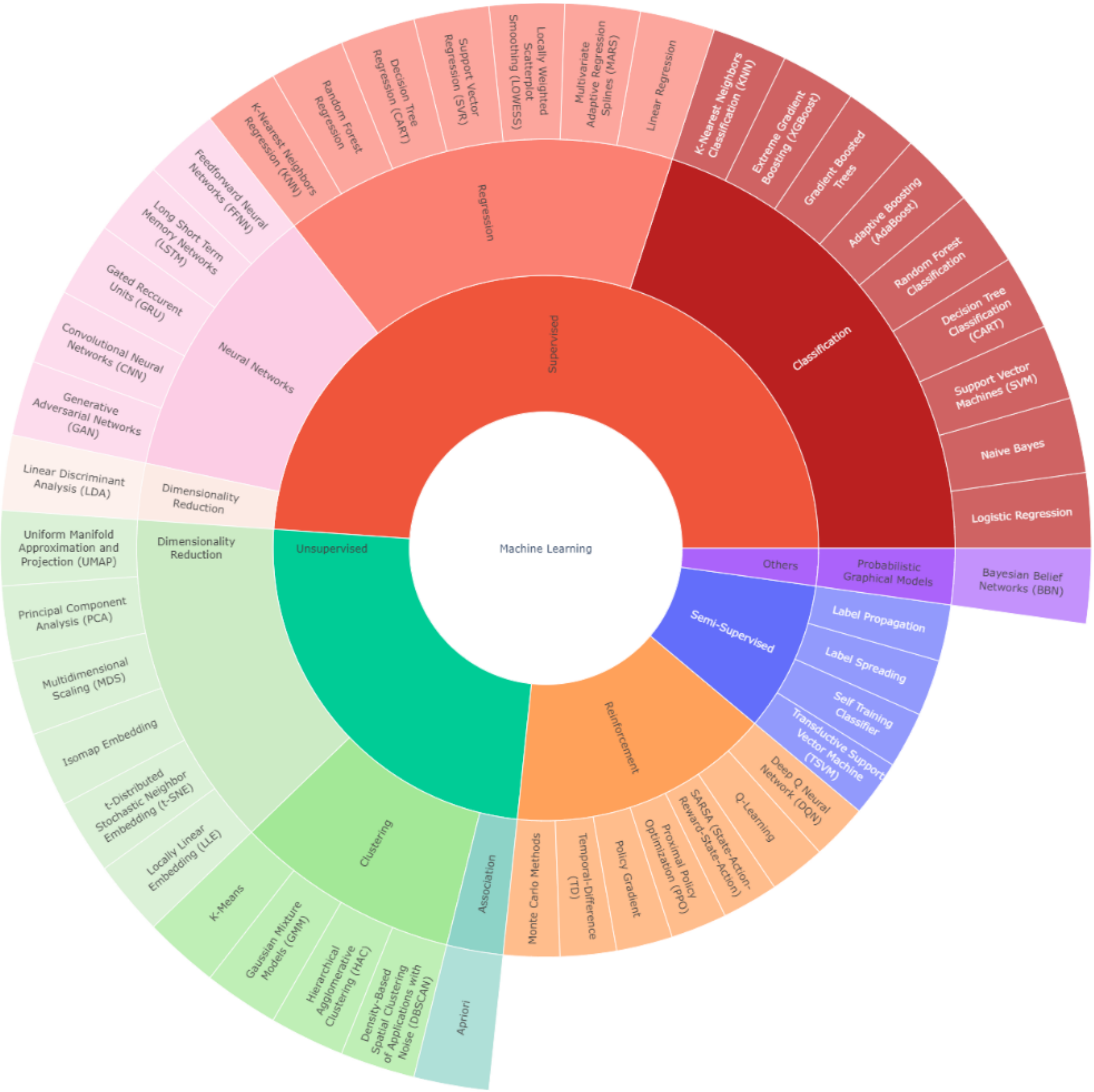

Machine Learning
Supervised
Regression
K-Nearest Neighbors Regression (KNN)
Random Forest Regression
Decision Tree Regression (CART)
Support Vector Regression (SVR)
Locally Weighted Scatterplot Smoothing (LOWESS)
Multivariate Adaptive Regression Splines (MARS)
Linear Regression
Classification
K-Nearest Neighbors Classification (KNN)
Extreme Gradient Boosting (XGBoost)
Gradient Boosted Trees
Adaptive Boosting (AdaBoost)
Random Forest Classification
Decision Tree Classification (CART)
Support Vector Machines (SVM)
Naive Bayes
Logistic Regression
Neural Networks
Feedforward Neural Networks (FFNN)
Long Short Term Memory Networks (LSTM)
Gated Reccurent Units (GRU)
Convolutional Neural Networks (CNN)
Generative Adversarial Networks (GAN)
Dimensionality Reduction
Linear Discriminant Analysis (LDA)
Others
Probabilistic Graphical Models
Bayesian Belief Networks (BBN)
Semi-Supervised
Label Propagation
Label Spreading
Self Training Classifier
Transductive Support Vector Machine (TSVM)
Reinforcement
Deep Q Neural Network (DQN)
Q-Learning
SARSA (State-Action-Reward-State-Action)
Proximal Policy Optimization (PPO)
Policy Gradient
Temporal-Difference (TD)
Monte Carlo Methods
Unsupervised
Dimensionality Reduction
Uniform Manifold Approximation and Projection (UMAP)
Principal Component Analysis (PCA)
Multidimensional Scaling (MDS)
Isomap Embedding
t-Distributed Stochastic Neighbor Embedding (t-SNE)
Locally Linear Embedding (LLE)
Clustering
K-Means
Gaussian Mixture Models (GMM)
Hierarchical Agglomerative Clustering (HAC)
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Association
Apriori

*Figure 6: General overview on machine learning methods and algorithms (Source: TowardsDataScience; 2022).*

*Table 2: Overview on the implementation of AI/ML techniques in various EO domains., including project examples, based on Salcedo-Sanz et al., 2020; Ma et al., 2019; Ferreira et al., 2020 and Reichstein et al., 2019. More information on ML – applications can be found in Ali et al., 2015.*

| **EO/remote sensing domain (processes)** | **Example** | **ML technique (non-limitative list)** |
|---|---|---|
| Image (data) fusion | Fusion of regional and local information (Yang *et al.*, 2019), fusion of evapotranspiration data retrieved from multiple satellite platforms (Knipper *et a*l., 2019), fusion of social media, RS data and topographic information (Rosser *et al.*, 2017). | DL, CNN (Ma *et al.*, 2019), SVM, Fuzzy C-means, AE, NN, CNN, RF, boosted algorithms, Kalman filter, SVR (Salcedo-Sanz *et al.*, 2020) |
| Image (scene) classification | Land use and land cover classification (Ma *et al.*, 2019), classification of crops (Wang *et al.*, 2019) | ANN, DL (e.g. CNN) (3) SVM, RF, NN (Ferreira *et al.*, 2020 + 3), decision tree (Ferreira *et al.*, 2020) |
| Image registration | Geo-localization accuracy improvement for optical satellite images (Merkle *et al.*, 2017), identifying corresponding patches in SAR and optical images with CNN (Hughes *et al.*, 2018a) | DL algorithms (Ma *et al.*, 2019) |
| Object detection | Object detection optical RS (Cheng & Han, 2016), CNN models for object detection in RS images (Ding *et al.*, 2018) | DL algorithms (e.g. CNN) (Ma *et al.*, 2019) |
| (Causal) pattern/ anomaly detection | Pattern detection by machine learning to detect smoke contamination in vineyards (Fuentes *et al.*, 2019), Relation flash flood LCLU by ML (Costache *et al.*, 2020) | DL algorithms (e.g. CNN) (Ma *et al.*, 2019; Reichstein *et al.*, 2019) |

| | | |
|---|---|---|
| Semantic segmentation | Semantic segmentation of large-scale 3D scenes and the extraction of building footprints and -heights (Ma *et al.*, 2019) | DL algorithms (e.g. CNN) (Ma *et al.*, 2019) |
| Downscaling | Drought monitoring by using downscaled (RF) high resolution soil moisture data (Park *et al.*, 2017) | CNN (Reichstein *et al.*, 2019) |
| Regression techniques | Global gridded soil information based on machine learning (Hengl *et al.*, 2017), Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques (Verrelst *et al.*, 2012) | RF (Reichstein *et al.*, 2019) |
| Change detection | Land Cover change detection (Zerrouki *et al.*, 2019), Change Detection Based on Machine Learning for Newly Constructed Building Areas (Wang *et al.*, 2021). Monitoring of dam reservoir storage (Sorkhabi *et al.*, 2022) | RF (Ma *et al.*, 2019) |
| Information prediction/reconstruction | Agricultural yield prediction (Yuan *et al.*, 2020), potato yields (Akhand *et al.*, 2016) & crop yields prediction (Bose *et al.*, 2016) | NN, CNN (Yuan *et al.*, 2020) |
| Parameter retrieval | Vegetation parameter retrieval (Yuan *et al.*, 2020), estimation of biomass (Jin *et al.*, 2019), land surface and air temperature (Tan *et al.*, 2019), soil moisture (Yuan *et al.*, 2020). Retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters (Pahlevan *et al.*, 2020). A chlorophyll-a algorithm for Landsat-8 based (Smith *et al.*, 2021). | DL algorithms, ANN (Ma *et al.*, 2019), mixture density networks (Smith *et al.*, 2021) |
| Time series analysis | Crop type classification using Landsat timeseries and ML techniques (Cai *et al.*, 2018) | DL algorithms (Ma *et al.*, 2019) |
| Compression of artifact reduction (Less common) | Compression artifacts reduction in remote sensing (Zhang *et al.*, 2018) | DL algorithms (Ma *et al.*, 2019) |

| | | |
|---|---|---|
| Network media data analysis (Less common) | New Perspectives to Improve Remote Sensing for Emergency Response, using social media (Li *et al.*, 2017) | DL algorithms (Ma *et al.*, 2019) |

*Table 3: Overview on relevant ML algorithms presently used in RS (EO) based on Ferreira et al. (2020), Holloway & Mengersen (2018) and TowardsDataScience (2022). Strengths and limitations are provided (expert knowledge).*

| Supervised | Description | Strengths | Limitations |
|---|---|---|---|
| **Classification algorithm** | | | |
| Support Vector Machines (SVM) | Algorithm used for linear and non-linear classification problems.<br>The separation of classes is denoted by a hyperplane in the transformed feature space. This hyperplane is optimized for separating the classes by the largest margin. The transformation of the feature space is performed with a kernel trick. | • Often a powerful method in specific use cases.<br>• Can include interactions when using nonlinear kernels.<br>• When using a linear kernel and optimized libraries a large amount of variables can be used. | • When using non-linear kernels and many variables the algorithm becomes intractably slow.<br>• To some extent a ’black box’ model. |
| Classification Trees | A classification tree is a structural mapping of binary decisions that lead to a decision about the class (interpretation) of a datapoint (observation).<br>The binary decisions are performed recursively by iterating over the input features. Values of the input features with the largest impurity decrease are chosen as a splitting criterion of the data. | • A very interpretable machine learning model.<br>• Implementation is fast.<br>• Can incorporate higher order interactions. | • Typically not a very well performing ML method because of its poor generalization. |

| | | | |
|---|---|---|---|
| Random Forest (RF) | Random Forest is an algorithm using several individual decision trees that work together as an ensemble. The prediction of the ensemble is more accurate than the predictions of the individual trees. While some individual trees provide a wrong prediction, the others (majority) provide a correct prediction. On average the ensemble will be still correct. Each individual tree makes use of input features that are randomly sampled out of the dataset. This results in unique decision trees.<br>There are some ML technique variations which resemble RF since they are also tree ensemble techniques. Among these are Bagging and Boosting. | • A powerful technique that often is among the best performing ML techniques 'out of the box'.<br>• Often provides already good results without performing extensive hyperparameter tuning.<br>• This technique is able to model higher order interactions.<br>• Can be used even with large amount of variables. | • Like most ML techniques this needs a lot of training samples.<br>• There can be some biases when dealing with classification.<br>• Rather slow training.<br>• To some extent 'black box' model. |
| K-Nearest Neighbour (K-NN) | This algorithm is based on the assumption that similar objects (data points) exist in close proximity of each other (small distance). The data point is classified in the same group as its nearest neighbors. | • Easy to implement and understand | • Compared to other ML techniques such as SVM, RF and XGB performances are on the lower end.<br>• It becomes very slow with larger datasets. |

| | | | |
|---|---|---|---|
| Naïve Bayes | Algorithm based on the conditional probability by the Bayes theorem, but assuming conditional independence between all input features. | • A rather fast algorithm.<br>• Often used to get a base accuracy.<br>• When the assumption hold this technique can perform better than logistic regression with less training samples. | • The assumption of conditional independence does not often hold.<br>• Compared to other ML techniques such as SVM, RF and XGB performances are on the lower end.<br>• Compared to logistic regression the output scores are not realistic probabilities. |
| Logistic regression | A classical statistical regression technique that can also be interpreted as an ML technique. This works like a linear regression but for the purpose of modelling categorical target variables. This technique models the target probability by making use of the logit link (sigmoid activation). The optimization and inference is performed by making use of the maximum likelihood estimation (MLE). Further extensions for regularization (L1, L2) can also be combined with logistic regression. | • Due to the statistical inference this technique has interpretable coefficients, confidence bands and target probabilities.<br>• Can be very powerful with correct feature engineering. | • Typically not the best performing ML model for complex patterns since it does not include higher order and non-linear interactions. |
| Maximum Likelihood Classification | Algorithm to predict the class of object X based on a probability distribution, given an observed dataset. This method is based on the assumption that the | • A prior distribution can be defined. | • A high number of samples for each class is required and often less |

| | | | |
|---|---|---|---|
| | image DN (digital number) values in each of the user-defined classes follow a multivariate normal probability distribution. This technique is often used in EO pixel analysis. | • A standard broadly applied algorithm available in GIS software packages. | accurate results are obtained than with full fletched ML techniques such as SVM and RF (eg. Volke & Abarca-Del-Rio 2020). The assumption of multivariate normal probability does not often hold. |
| Neural networks (NN) | A model that consists of several regressors (neurons) with trainable weights and activation functions. These are stacked in layers. The output from one layer serves as input for the other. The whole superposition of layers can be optimized to learn to predict a certain target variable. This optimization is performed in batches and with gradient descent. When many layers (e.g.: more than 3) are stacked we refer to these neural networks as 'deep learning'.<br>A popular type of deep learning architecture often used for the classification of images and within EO is the convolutional neural network.<br>In classification a sigmoïd or softmax activation function is used in the last layer to emulate a probability estimate. | • When enough data is present this technique is among the best performing ML algorithms.<br>• Very large, out of memory, datasets can be used since the optimization is performed in batches.<br>• Higher dimensional input data can be used. Therefore these ML algorithms can train on data like images and videos.<br>• Complex learning systems can be engineered. | • Large amounts of data is necessary. These large annotated datasets are rare.<br>• Since most impressive results are achieved at large scales (large datasets and large models) setting up a correct system architecture can be challenging.<br>• With deep learning long training times and lots of architectural and hyperparameter tuning is necessary. Therefore |

| | | | |
|---|---|---|---|
| | | • There are several pretrained networks available. These can be used for transfer learning such that less data is necessary compared to training a model from scratch. | experimentation can be very time consuming.<br>• Up till now the deep learning models are still ’black box’. |
| Extreme Gradient Boosting (XGB) | Just like random forest the XGB is a tree ensemble method. However the trees are not trained in parallel but are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting. The values of the splitting criterion within the trees are found by gradient descent. This boosts the speed of learning. | • A state of the art technique which is often one of the best performing in any field it gets applied.<br>• Achieves good results ’out of the box’.<br>• Has optimized libraries in python which include fast SHAP calculations for model explanation.<br>• An arbitrary loss function can be chosen.<br>• Can handle large amounts of input variables. | • Even with the SHAP explain ability this model is still to some extend ’black box’.<br>• No higher dimensional data like in NN. |

| | | | |
|---|---|---|---|
| **Supervised/dimension reduction algorithm** | | | |
| (Linear) Discriminant Analysis | This method finds a linear combination of features that characterizes or separates two or more classes of the observations. The dimension reduction obtained can be used for visualizing the separation of classes in a lower dimensional space. | • Can be used for insights into the data due to dimension reduction. | • Often not a very powerful ML technique. |
| **Regression algorithm** | | | |
| Linear regression | A classical statistical technique. It is a method for modelling the linear relationship between dependent and independent variables. | • Due to the classical statistical inference framework, robust coefficient estimates and confidence/prediction bands can be obtained. This leads to a highly explainable model. | • Often not the best performing regression method. |
| Regularized regression | To avoid overfitting and multicollinearity an inherent regularization can be performed. This decreases the model information for better generalization. Common types of regularization in regression are L1 and L2. L1 regularization can be used as a feature selection method. Regression  with L1 regularization | • Inherent regularization can be useful when we do not want to perform the feature selection ourselves, hence it | • Similar as in regression |

| | | | |
|---|---|---|---|
| | is called Lasso. Regression with L2 is called Ridge regression. When the two types are combined it is called an elastic net. | can be more automated than regression. | |
| Geographically weighted regression | A classical linear regression has an assumption of stationarity. This model incorporates varying coefficients depending on geolocation. | • Can give better results when used in a geospatial context. | • Current implementations in python can handle limited dataset sizes. |
| Regression Trees (RT's) | Identical to a classification tree with the distinction that the outcome value is a real number, not a class (e.g. price of a house). | • Similar as in classification | • Similar as in classification |
| Support Vector Regression (SVR) | Based on the SVM principle, however, the best fit is the hyperplane containing the maximum data points possible. | • Similar as in classification | • Similar as in classification |
| Neural Networks (NN) | Similar as in classification. The last layer however does not have a sigmoïd/softmax activation but a linear activation. | • Similar as in classification | • Similar as in classification |
| Gaussian Process Regression (GPR) | Probabilistic model (nonparametric, kernel based) allowing to make predictions whilst providing uncertainty measurements for these predictions. | • Provides nonparametric prediction bands. | • Becomes intractable on large datasets. |
| Kernel Ridge Regression | Here a Ridge regression is combined with the kernel trick to obtain a regression in a transformed feature space, just like SVR. | • Training kernel ridge regression is faster than SVR on medium sized | • SVR scales better for large datasets. |

| | | | |
|---|---|---|---|
| | | datasets (less than 1000 samples). | |
| Partial Least Square Regression | A regression method with inherent dimension reduction, just like PCA. This dimension reduction can serve as a regularization. | • Similar as in classification | • Similar as in classification |
| Random Forest | The regression version is based on regression trees. | • Similar as in classification | • Similar as in classification |
| Extreme gradient boosting | The regression version is based on regression trees. | • Similar as in classification | • Similar as in classification |

| **Unsupervised** | **Description** | **Strengths** | **Limitations** |
|---|---|---|---|
| **Clustering algorithm** | | | |
| Gaussian Mixture Models | A generalization to K-means by providing a probability for each data point to belong to a certain cluster based on the covariance structure of the data and assuming a multivariate gaussian distribution of the clusters. | • Probability estimates. | • This algorithm requires the number of clusters to be specified. |
| K-Means | The K-Means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. | • It scales well to large number of samples and has been used across a large range of application | • This algorithm requires the number of clusters to be specified. |

| | | | |
|---|---|---|---|
| | | areas in many different fields. | |
| **Clustering/dimension reduction algorithm** | | | |
| Self organising map | Technique to rescale a high-dimensional dataset (ρ variables, n observations) to a 2-dimensional representation. The technique makes use of ”nodes” that are randomly placed in the dataset. An observation $i$ of the dataset ($n$) will select its closest node after which the node will move to its direction (neighboring nodes will be move as well but in a lesser extent, this is not the case for K-Means). This is an iterative process taking place for each observation $i$ of $n$. This iterative process results in the nodes being the center of the thereby defined clusters. Topological relation of the original dataset is preserved. | • The output provides a clear insight.<br>• No number of clusters to be defined in advance. | • A lot of hyperparameters have to be defined in advance. |
| Density based Spatial Clustering of Application with Noise (DBSCAN) | Clustering technique using a minimum distance criteria and a minimum density of neighboring points (within this predefined distance) approach. Technique that identifies outliers and doesn’t incorporate them in a cluster. | • It does not require one to specify the number of clusters in the data a priori, as opposed to k-means. | • It fails in case of varying density clusters.<br>• It does not work well in case of high dimensional data. |

| | | | |
|---|---|---|---|
| | | • It can find arbitrarily-shaped clusters<br>• It has a notion of noise, is robust to outliers and can detect outliers | |
| Hierarchical Cluster Analysis | Algorithm to classify similar objects in a non predefinednumber of clusters, each object being its own cluster at the start (bottom-up/agglomerative approach), to be merged with another similar cluster in the next step. A hierarchy clustering is build this way. Clustering can also be top-down (divisive). | • The advantage of hierarchical clustering is that it is easy to understand and implement. The dendrogram output of the algorithm can be used to understand the big picture as well as the groups in your data. | • The weaknesses are that it rarely provides the best solution, it involves lots of arbitrary decisions, it does not work with missing data, it works poorly with mixed data types, it does not work well on very large data sets, and its main output, the dendrogram, is commonly misinterpreted. |
| Fuzzy C-Means (FCM) | Clustering technique (similar to K-Means) by which a datapoint can belong to more than 1 cluster (based on degrees of membership). | • It gives the flexibility to express that data points can belong to more than one cluster. | • A-priori specification of the number of clusters. With lower value of β we get the better result but at the expense of more number of iteration. Euclidean distance measures can unequally weight underlying factors. |

| Dimension Reduction | | | |
|---|---|---|---|
| Principal component analysis (PCA) | Technique that reduces the amount of features in a dataset (lower dimensionality) making it more easy to analyze and visualize, however losing some information. | • Benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data. | • Low interpretability of principal components.<br>• The trade-off between information loss and dimensionality reduction. |
| T-Distributed Stochastic Neighbour Embedding | Method for visualizing high dimensional data in a low dimensional space (2D/3D). It is a non-linear dimensionality reduction technique. | • Handles Non Linear Data Efficiently<br>• Preserves Local and Global Structure | • Computationally Complex<br>• Non-deterministic<br>• Requires Hyperparameter Tuning<br>• Noisy Patterns |
| Isomap | Non-linear dimension reduction technique used for low dimensional embedding. | • Handles Non Linear Data Efficiently<br>• Preserves "true" relationship between data points | • Computationally expensive |
| Autoencoder | An autoencoder is neural network architecture with a decoder architecture on top of a encoder | • Autoencoders can be a powerful method to | • Apart from the fact that large annotated datasets |

| | | | |
|---|---|---|---|
| | architecture. This autoencoder is trained on data to predict itself after encoding. Hence the encoder can be used to produce a lower dimensional embedding in which the optimal information is retained | pretrain neural networks on unsupervised data. Unsupervised datasets are typically very large datasets and hence very large models can be trained for optimally encoding information. The encoder can be decoupled and used for a supervised training task which hence require much less annotated data.<br>• Since the technique involves neural networks it can be used on datatypes with higher dimensions.<br>• They are often used as part of larger AI systems such as GAN’s | are not needed, similar issues arise as in neural networks in classification |

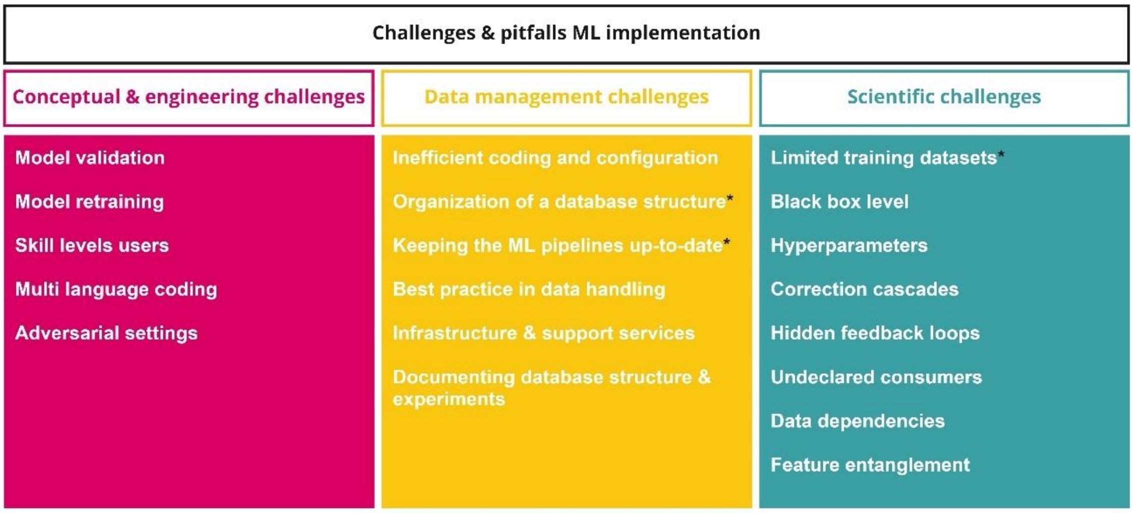
## 4.2 Challenges & pitfalls of the implementation of ML

The use of ML systems for EO purposes has been increasing rapidly the past decade. Although a lot of ML algorithms and ML based applications are currently in operation by research institutes or the industry, different challenges with regard to ML systems have been identified.

Scientific papers (Jentzsch *et al.*, 2021; Nascimento *et al.*, 2019; Sculley *et al.*, 2014; Shelter *et al.*, 2018) mention a variety on challenges in ML implementation, however they can be generally grouped as:

- ***Conceptual challenges***: challenges on the vision of how a ML model should look like and how to handle a model (model validation, model retraining, …)
- ***Data management challenges***: challenges in structuring and handling the data in an effective and long-term sustainable way
- ***Scientific challenges***: pitfalls and difficulties encountered processing the data

Figure 7 provides an overview on the challenges identified for each group, based on the works of Jentzsch *et al.* (2021), Nascimento *et al.* (2019), Sculley *et al.* (2014) and Shelter *et al.* (2018).



**Challenges & pitfalls ML implementation**

| Conceptual & engineering challenges | Data management challenges | Scientific challenges |
|---|---|---|
| Model validation | Inefficient coding and configuration | Limited training datasets* |
| Model retraining | Organization of a database structure* | Black box level |
| Skill levels users | Keeping the ML pipelines up-to-date* | Hyperparameters |
| Multi language coding | Best practice in data handling | Correction cascades |
| Adversarial settings | Infrastructure & support services | Hidden feedback loops |
| | Documenting database structure & experiments | Undeclared consumers |
| | | Data dependencies |
| | | Feature entanglement |

*Figure 7: Overview of ML challenges, conceptual- data management- and scientific challenges. * Indicates challenges stressed most in literature.*

***Model validation*** is considered a conceptual challenge by Shelter *et al.* (2018), since decisions on how to revalidate the model, how to define the train/validation/test split and backtesting are not always straightforward. How to define the train/validation/test split all depends on the scenario, e.g. a forecasting situation will have a different split then other scenarios. And although backtesting of the models over time is necessary to validate their accuracy (due to changes in data, code, software dependency, …), it implies the models should have been trained on the same training, testing and validation dataset, using the exact same code. ***Retraining of models*** occurs when new data events take place in the dataset. Shelter *et al.* (2018) gives the example of a new public holiday or promotional activity in the retail demand forecasting domain. This new event can have a serious impact on the accuracy of the model. However, managing evaluation and training data over time, when ML pipelines are continuously changing is a challenge (Shelter *et al.*, 2018).

***Multi language code*** is considered a significant engineering challenge since it is hard to keep all components consistent and to perform error checks across the language barrier. Such code bases are difficult to handle later onwards since different component setups have to be orchestrated to work as one (Shelter *et al.*, 2018). Different ***skill levels*** of users might pose problems as well, in certain cases leading to suboptimal ML modelling (Shelter *et al.*, 2018).

The biggest challenge stressed in literature, concerning data management, is the ***efficient organization of a database structure***. A lot of effort needs to be spent to structuring the data, creating a solid base for the model (Nascimento *et al.*, 2019). Experts indicate that organizing ML projects and experiments from scratch is very hard, maintaining a clear and comprehensive overview being even more difficult when the system keeps growing and/or multiple developers are involved (Jentzsch *et al.*, 2021; Nascimento *et al.*, 2019). Since every minor change e.g. in the code, results in a new experiment, managing all these experiments and be able to keep track of the metadata, workflows and lineage of models/different ML pipelines involved is a major challenge (Jentzsch *et al.*, 2021; Nascimento *et al.*, 2019; Sculley *et al.*, 2014).

More specifically, ***ML pipelines*** most often comprise different operations (data integration, feature transformation, model training), each based on different abstractions. In complex analytic problems, dataflow abstraction is often very limited, since models are implemented as black boxes. It is thereby very difficult to abstract metadata from a specific pipeline (Shelter et al., 2018). Experts pointed out is hard to keep track which adaptations were made in which experiment leading to which result. Furthermore, the high frequency of updates of tools and programming packages increases the effort needed to keep the ML pipelines up-to-date and aligned (Jentzsch *et al.*, 2021).

Experts also indicate resources to thoroughly document the experiments and workflows/metadata are often scarce. ***Tools to organize experiments*** and answer metadata related questions (e.g. what part of the data set was used?) exist, though their feasibility is questioned by the experts (Jentzsch *et al.*, 2021).

***Documentation on "how to handle the data"*** is only very limited available. There is no best practice and no real standardized processes on what type of tasks should be performed in this phase, e.g. checking missing data, verifying inconsistencies, thresholds to use, ... No real guidelines are provided, resulting in every expert deciding on how to proceed in his/her own manner (Jentzsch *et al.*, 2021; Nascimento *et al.*, 2019)

Concerning infrastructure, ***high transfer costs*** were mentioned to transfer existing code from one platform to another (e.g. Caffe - TensorFlow) (Jentzsch *et al.*, 2021).

Finally ***inefficient coding***, originating from "glue code", deficiencies in ML system configuration options or the presence of "dead" experimental codepaths ("experimental code as a conditional branch within the main production code") poses even more challenges (Sculley *et al.*, 2014).

A third category of challenges comprises the "scientific" challenges, the most important challenge being the availability of ***training datasets*** (labeled or unlabeled). In various scientific papers the importance of these datasets is stressed, their limited availability being indicated as a significant obstacle (Ma *et al.*, 2019; Reichstein *et al.*, 2019; Sagan *et al.*, 2020).

ML systems are also often referred to as ***black boxes*** since it is difficult to get a grip on the processes taking place between input and output data in a system, e.g. the opacity/interpretability of a DNN as explained by Jentzch *et al.* (2021) and Reichstein *et al.* (2019).

***Other challenges*** are the impact of hyperparameters, correction cascades, undeclared consumers, hidden feedback loops, data dependencies, feature entanglement and prediction uncertainties. More information on these specific challenges can be found in Jentzch *et al.* (2021), Reichstein *et al.* (2019) and Sculley *et al* (2014). Ultimately, predictions of deep learning models, even though the model has a high accuracy, might be implausible due to the presence of observational biases in the data and/or extrapolations. Therefore, domain specific knowledge should be integrated in the model to establish theoretical constraints (rules on physics of the Earth System) (Reichstein *et al.*, 2019).

# 5 Implementation of AI for optimal exploitation of EO data

This section discusses the ways in which AI can augment the exploitation of EO data and possibly create innovative approaches to support or solve EO end user needs. It results from a 2-step work process in which first end user needs were identified and secondly, the ways in which AI can be of support to the EO community and its user needs analyzed. Since terminology such as "end user" and "user needs/requirements" will be frequently used in the consecutive sections, the following definitions are provided:

- **End user**: in this deliverable an end user can be part of the wider base of businesses, institutional players, consumers, industries, research institutes, governments, and nonprofit organizations.  However, end users are experts in either EO, AI or both and are actively working in the field of data-processing, -assimilation, -management or application building (EO/AI related).

- **User needs and requirements**: User needs and requirements describe any function, constraint, specification, observation, wish list, service or other property that must be provided to satisfy the current and future needs of the user. They can also relate to potential existing gaps between users' aims and their current situation, which is reflected by user difficulties and opportunities, as well as the context of use, which comprises the intended users' attributes, current tasks, and environment.[1] User requirements are created from the perspective of the user. Any function, constraint, or other property that must be given to meet the user's needs is referred to as a user need[2] (Water-ForCE, 2022a).

[1] Kujala, Sari et al. "Bridging the Gap between User Needs and User Requirements." (2001).

[2] Abbott, R. J. An Integrated Approach to Software Development. Wiley, New York, 1986.

## 5.1 End user needs of the EO community

Based on the output deliverables of WP's 1 to 5 and the insights obtained from several Water-ForCE workshops, an extensive list of end user needs and -requirements was created. Of this list, a selection was made, considering 4 user need categories for which AI might be able to offer new techniques or innovative approaches. The selection was also based on the weight of the end user need (how many people indicated the problem) and the potential impact an "AI solution" of the user need would have on the EO and modelling community. It concerns user needs related to data (pre-) processing, data characteristics (e.g. spatial resolution) and data retrieval.

Table 4 gives an overview on the selected end user needs and what they encompass. In the following sub-sections each of these topics is discussed in detail, describing observed difficulties and providing examples.

*Table 4: Overview on selected end user needs.*

| User need category | Description |
|---|---|
| **Pre-processing** | |
| - **Optical** | The pre-processing of optical remotely sensed data includes a variety of steps, depending on the data that is used. However, considering the existing needs, especially enhanced algorithms for *atmospheric corrections (atmospheric absorption, sky-glint, sun glint,...)* are in high demand and therefore listed here as an important user need. |
| - **Radar** | The EO community indicates the complexity of SAR backscattered signals and how it can quickly become overwhelming. Pre-processing steps are considered highly complex and although there is a high potential for implementation of AI more extensive research is needed in this field. |

| | |
|---|---|
| **Resolution** | Different or limited *spatial, temporal and spectral resolutions* are often the limiting factor for the utilization of satellite image data for different applications and therefore often indicated as an important user need. Enhancement of these resolutions may lead to higher quality products. |
| **Parameter retrieval** | The need for a *higher variety in parameter products* (e.g. more groundwater products) was indicated as an important user need by the EO community. ML approaches are already of significant importance in the field of parameter retrieval and believed to have a high potential impact when further developed. |
| **Image classification** | Image classification plays an important role in exploiting EO data at all levels. However, due to the characteristics of RS data such as high dimensionality and relatively small amounts of labeled samples available, performing RS image classification faces great scientific and practical challenges. Literature and experts indicated the potential role of machine learning in tackling those issues. |

### 5.1.1 Pre-processing

#### Optical

The pre-processing of optical remotely sensed data includes a variety of steps that depend on the data that is used. For example, for Sentinel-2, at Level-1B, the OLCI instrument provides radiance measurements of the Earth's surface in the visible and near infra-red spectrum. These measurements are accurately calibrated Top-Of-Atmosphere (TOA) radiances, annotated with geo-referencing data and observation geometry parameters. The first pre-processing step aims to convert Level-1 TOA radiance into reflectance values that are relevant for geophysical properties. This step is followed by e.g. corrections for gaseous absorption (Atmospheric Correction), pixel classification (in particular, water/land/cloud) and retrieval of the total column water vapor content. This deliverable focuses on the user need Atmospheric corrections.

***Atmospheric correction (AC)*** is the process of compensating for atmospheric scattering and absorption.and for surface reflection at the air-water interface (i.e., sky-glint and sun-glint) from the signal measured at the Top of Atmosphere (TOA). This compensation is essential for the accurate retrieval of aquatic reflectance and downstream science products (e.g., near-surface concentration of chlorophyll-a (Chla), and Total Suspended Solids (TSS)). AC over open water is carried out adequately since quite some time as mentioned by the International Ocean Color Coordinating Group (IOCCG) in 2010. AC over coastal and inland waters however still leads to large uncertainties in derived satellite data products.

AC algorithms are situated largely in two categories (Pahlevan et al., 2021):

- A two-step process where the effects of Rayleigh and gaseous absorption are first removed and then aerosol contribution is approximated and,
- Machine Learning techniques.

A comparison was done by Warren *et al.* (2019) of six commonly used AC algorithms for Sentinel2 (Acolite, C2RCC, iCOR, l2gen (also known as SeaDAS), Polymer and Sen2Cor). C2RCC is based on ML, the others are two-step techniques. Although all are considered mature, all still showed very high levels of uncertainties and low $R_2$ against in situ datasets above coastal and inland waters. Especially the Red and NIR bands show poor results which is of concern for inland water monitoring. These bands are paramount to the determination of key water parameters.

Another point of attention is that multispectral images can be affected by adjacency effects up to 20km inland of a shoreline (Pan *et al.*, 2022; Warren *et al.*, 2019). Out of the six algorithms mentioned above, only iCOR has a built-in adjacency effect corrector. Additionally, according to (ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters - ScienceDirect) AC algorithms should incorporate more representative aerosol types and/or bio-optical models depending on the underlying mechanisms of the AC processors.

### Radar

Rapid-revisit Synthetic Aperture Radar (SAR) satellites in Low Earth Orbit (LEO) are promising thanks to their observational persistence, low cost and their ability to “see in the night”.  But the complexity of the backscattered signal can quickly become overwhelming and prevent data-scientists to include SAR data in their analyses. Pre-processing the data is the most complex part of the work. Depending on the research question at hand, pre-processing can include steps like applying an orbit file, radiometric calibration, multilooking, de-bursting, reducing speckle noise and terrain correction (Meyer).

Preprocessing techniques achieve the effects of suppressing background redundancy and enhancing target characteristics by processing the size and gray distribution of the original SAR image, thereby improving the downstream application. Speckle noise for example is a result of the interference of many waves of the same frequency, having different phases and amplitudes, which add together to give a resultant wave whose amplitude, and therefore intensity, varies randomly. This causes blurring and leads to loss of the information of the objects.

### 5.1.2 Resolution

Different and limited spatial, temporal and spectral resolutions are the limiting factor for the utilization of the satellite image data for different applications. Unfortunately, because of technical constraints, satellite remote sensing systems are faced with a tradeoff between the resolution types. E.g., constellations that offer high spectral resolution are often faced with medium or low spatial resolutions. Image fusion, harmonization and super-resolution are just some examples of techniques that aim to overcome the trade-off issue.

Additionally, the concept of Virtual Constellations (VC) is gaining interest within the science community owing to the increasing number of satellites/sensors in operation with similar characteristics. Sen2Like, for example, offers a solution for harmonizing and fusing Landsat 8/Landsat 9 data with Sentinel-2 data. Sen2Like processes a large collection of Level 1/Level 2A products and generates high quality Level 2 Analysis Ready Data (ARD) as part of harmonized (Level 2H) and/or fused (Level 2F) products providing high temporal resolutions (Saunier *et al.*, 2022). In the era of artificial intelligence, S2L also has the potential to improve the model training processes by providing radiometrically consistent spatiotemporal information and might be helpful for many applications in the field.

### 5.1.3 Parameter retrieval

Machine learning has been widely used as a powerful tool for groundwater and surface water applications.  It can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Many approaches have been tested: SVM, CNN, DNN, PLS but several challenges remain in fully applying machine learning approaches in this field. Often it is not a one size fits all situation. Applying specific algorithms that respond to the peculiarities of a region can deliver better results. E.g. for surface chlorophyll concentration in the Black Sea, the Ocean Colour Thematic Assembly Center of the Copernicus Marine Service (CMEMS*) use a neural network closer to the coast and an analytical algorithm in the open ocean. Moving further from the coast, the neural network gets further from its comfort range and the analytical algorithm becomes more useful.

Machine learning is usually dependent on large amounts of high-quality data. Obtaining sufficient data with high accuracy in water treatment and management systems is often difficult owing to the cost or technology limitations. Furthermore, the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hinders the wide application of machine learning approaches. The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.

To overcome the above-mentioned challenges, this field needs more advanced sensors applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches. Also the feasibility and reliability of the algorithms should be improved. Lastly it asks for more interdisciplinary talent with knowledge in different fields (both EO and AI) to develop more advanced machine learning techniques and apply them in engineering practices.

### 5.1.4 Image classification

Image classification plays an important role in exploiting EO data at all levels, going from water/land/cloud detection to object recognition and creating Land-Use-Land-Cover (LULC) maps. However, due to the characteristics of RS data such as high dimensionality and relatively small amounts of labeled samples available, performing RS image classification faces great scientific and practical challenges. Some state of the art technologies are discussed below.

- ***Object-based image analysis (OBIA)***: discipline devoted to partitioning remote sensing (RS) imagery into meaningful image-objects and assessing their characteristics through spatial, spectral and temporal scale (Ma *et al.*, 2019). Nowadays a patch-based strategy is a generally accepted method, which integrates CNNs with OBIA. The critical issue with this approach is how to determine the values of the relevant parameters (e.g., patch size), because classification accuracy is largely affected by these parameter values.

- ***Semantic segmentation***: Semantic segmentation aims to assign labels to each pixel in an image. Facilitated by deep CNNs, especially by the end-to-end fully convolutional network (FCN), interest in semantic segmentation of remote sensing images has increased in recent years. Still several issues need to be addressed. Among them are problems with imbalanced classes, trade-offs between downsampling and accurate boundary localization and overcoming the difficulty of high variety of objects, especially in high spatial resolution images. (Ma *et al.*, 2019)

- ***Scene classification/Object detection***: Applications that are frequently confused. Scene classification is a procedure to determine the image categories from numerous pictures. Object detection aims to detect different objects in a single image scene. Current studies prefer to extract certain specific type(s) of objects (airplanes, cars, etc.) from high-resolution images through a fixed window size, either through scene classification or object detection (Ma *et al*, 2019). However, more data and object types of objects are encountered in practical remote-sensing applications–for example, medium-resolution Landsat data and Sentinel data. Therefore, how to design the effective algorithms to overcome the difficulties emerging from different-scale objects (the different type of objects often appears at different scales in remote-sensing images, and also the same object can have variable size in different-scale remote sensing images) is an urgent problem in both subfields (Deng *et al.*, 2018).

## 5.2 The implementation of AI with regard to EO community needs

Understanding the needs of the EO community, the following sub-chapters try to provide an overview on current best practice and possible approaches to tackle those posing challenges in EO data-assimilation and -modelling.

### 5.2.1 Consultation of stakeholders and knowledge experts

Although a literature review was carried out, presenting a lot of different ML algorithms suited for various applications in EO domains (§4), also the EO and AI community was consulted on how AI can support in addressing end user needs and which specific AI/ML techniques can be used for specific problems. In order to get a clear view on how and for which type of applications/processes stakeholders use AI techniques, several initiatives were taken:

- ***Face-to-face meetings*** with stakeholders. Twenty-two companies were selected by their core business, all related to the use of AI/ML for research projects or product development. The meetings create the opportunity to have a discussion on how they apply AI within the company business and which techniques/algorithms they use. In general, how a company/institute relates to the use of earth observation data and/or the use of artificial intelligence for the exploitation of EO data. Response levels however were very low. Annex 1 provides an overview of the companies that were contacted.
    - ***Survey***: due to low response levels on the Face-to-face meetings a survey was sent to 5 additional companies/institutes (An overview on the companies and survey questions can be found in Annex 2).

- ***Workshop participation (1):*** in order to get feedback from the EO-AI community on the implementation of AI for atmospheric corrections (considered an important end user need), questions were posed to the public and speakers during the Water Quality Continuum Atmospheric Correction Workshop on (20 October 2022, ± 55 participants). Questions were e.g.:

  - AI is already used in several methodologies for AC. Are there still gaps? What are the gaps? What are recommendations?
  - Is sunglint still an issue? Do you think AI techniques can help to detect or correct sunglint?

- ***Workshop participation (2)***: Participation in the workshop "6th WGNE workshop on systematic errors in weather and climate models – part Machine learning/AI and data assimilation" to gather information on how AI is used for systematic errors at this moment.

- ***Water-ForCE Webinar***: In the webinar "Technical Needs for Copernicus Inland Water Monitoring Service" (26 October 2022, ± 35 participants) the goal of WP5.3 "Exploring the use of Artificial Intelligence (AI) to Optimize the Exploitation of satellite EO and modelling data" was introduced. A Mentimeter poll was organized to capture the audience feedback on the use of AI for EO. Questions and responses can be found in Annex 3.

Despite the attempts to collect feedback from the EO-AI community, response levels were in general (very) low. However, some conclusions can be made:

- AI/ML techniques are most commonly used for atmospheric corrections, regression analysis, clustering, parameter retrieval and identification of related parameters (based on the output of the poll (Figure 8), survey (Annex 2) and face-to-face meetings).
- Widely used AI/ML algorithms are (Figure 9):
  - Neural Networks (NN)
    - CNN ( e.g. for image classification, object detection, image segmentation)
    - FRCNN (for better downsampling, upsampling and "super resolution" creation)
  - C2RCC (e.g. atmospheric corrections)

  - Generative Adversial Networks (to enhance spatial resolution of satellite images)
- The end user needs identified in §5.1 (parameter retrieval, atmospheric corrections and enhanced resolution) are confirmed by the stakeholders and experts.
- The main limitation and/or bottleneck for the use of ML are the availability of labeled datasets and the capacity of finding the correct tools/choosing the correct technique. Also the explainability of the results proves to be an issue.
- ML offers an extensive amount of possible algorithms and techniques, however, there are no clear "guidelines" in how to choose the best technique as often the goal of the study is very different and several options are possible for the same problem. Initiatives for implementing AI in EO data assimilation and -processing are rising but information is scattered and users don't always succeed in accessing relevant data or information platforms.

The above mentioned implementation purposes, techniques and algorithms and limitations are validated by literature as stated in sections 4.1 and 4.2.

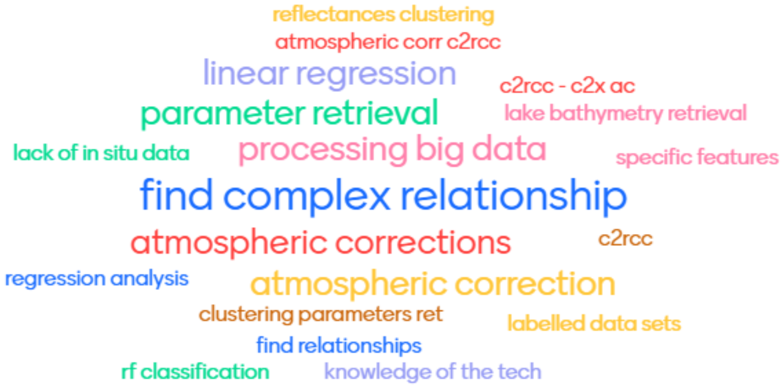
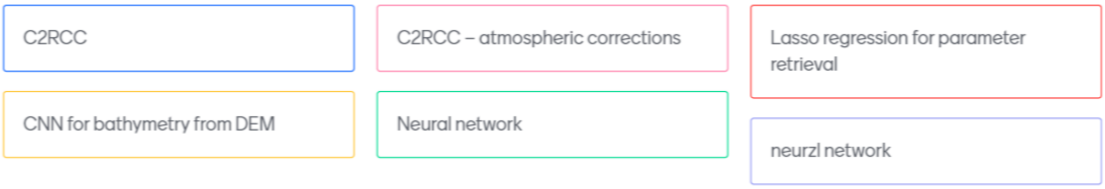


*Figure 8: Responses of the Mentimeter poll to the question "In which cases do you use AI/ML techniques".*



C2RCC

C2RCC – atmospheric corrections

Lasso regression for parameter retrieval

CNN for bathymetry from DEM

Neural network

neurzl network

*Figure 9: Responses of the Mentimeter poll to the question "Which AI/ML algorithms do you use most?".*

### 5.2.2 AI techniques and direct or first-order impacts versus indirect impacts

Pinpointing how AI could help accelerate the process of servicing the scientific community by developing data products that answer to their user needs is ambiguous. The impacts of AI techniques situate themselves from direct or first-order impacts to indirect or second or third order effects. Many examples can be given of which here are presented a few.

***Image registration*** is a method where two or more images, captured by different sensors, or at different times, are aligned. Image fusion entails the actual integration of images. Image registration is a step in which the source image is mapped with respect to the reference image (Viergever *et al.*, 2016). This is a fundamental task in several remote sensing tasks such as image fusion and change detection (Zhao *et al.*, 2021). Hence, it follows that this field could benefit from AI on multiple levels: from registration (1st) to fusion (2nd) and change detection (3d).

***Sunglint*** (user need AC) for example, significantly impacts the detection capacity of many remote sensing applications. While strong glint cannot be dealt with and affected pixels need to be excluded from further processing, weak wave glint can potentially be corrected and allow the extraction of meaningful information from the affected pixels. Many classical approaches to remove sunglint assume a negligible marine contribution in the NIR/SWIR which is invalid for turbid environments. (Giles *et al.*, 2021; Tapouzelis *et al.*, 2021) Moreover, some applications such as the detection of Floating Marine Litter (FML) use exactly these wavelengths. Correction of the signals therefore inevitably impact the performance of the detection methods (Tapouzelis *et al.* 2021). The future of using spectral remote sensing to detect floating marine litter is equally dependent on the pre-processing of images than on the used classifiers. Sunglint correction is one of the main image pre-processing steps. Hence enhancing sunglint correction methods by implementing ML could significantly affect the future accuracy of this field. In this case, implementing ML has a direct effect on the user need sunglint.

Successful ***parameter retrieval*** and ***object detection*** in coastal and inland waters is affected by adjacency effects and atmospheric correction. Meanwhile also atmospheric correction algorithms are affected by adjacency effects. Furthermore, it is often overlooked that AC algorithms that are based on ML need proper optical specification of the underlying water and atmosphere conditions (Brockman *et al.*, 2016; Schiller *et al.*, 1999). These algorithms are trained on synthetic datasets created by radiative transfer models. Hence it is clear that the successful training of these algorithms is largely dependent on the accurate outcomes of the RTMs and that improvements in the RTMs will lead to more reliable AC algorithms.

Such complex interacting effects make it difficult to compare over different case studies, different algorithms, different data sources. It could be argued that at least for some of the user needs identified, the appropriate answer should not be to try and advise a one-size-fits-all solution (see §5.2.5). For those examples it could be better to develop a useful tool that helps scientists to develop a workflow that is tailored to their research question(s) at hand.

### 5.2.3 Bottlenecks implementing AI

Although AI is a powerful tool, its implementation and reaching its full potential is limited by different factors. The most important causes are discussed in detail in the next paragraphs and are based on stakeholder and expert consultations.

**Benchmark datasets**

Only in comparison to existing knowledge can method performance be assessed. For that purpose, benchmark datasets with known and verified outcome are needed. High-quality benchmark datasets are valuable and may be difficult, laborious and time consuming to generate (Sarkar *et al.*, 2020).

A good benchmark dataset checks at least a few of the following boxes;

- Open/discoverable/accessible
- Has enough features
- Is labeled
- Is well documented
- Is accompanied by a demonstration (e.g. a script or notebook)

One of the main problems affecting large-scale remote sensing data processing is the lack of labeled samples. This becomes more evident when training deep learning models, which require a considerable number of labeled data to obtain good generalization capability. Collecting field data or manually creating labeled data is an operational burden. Moreover, also the quality and the representativeness of the training samples is important. It is interesting to explore more learning strategies for collecting labeled data in a fast and efficient way (e.g., active learning or transfer learning). Also cartographic products, thematically available on different scales represent a valuable source of information to generate large scale reference data.

Ocean Scan, for example, aims to provide an answer to the main challenge in training AI algorithms for the purpose of monitoring the water surface for the presence of plastics. In this area, the lack of in-situ ground truth data that allow to reliably label images is a bottleneck to advance in this field. Ocean Scan is a labeled database created to promote collaboration and research in the field of marine litter. It collects global in-situ observations of marine litter from whomever wants to contribute and associates them to their corresponding Earth Observation images from different missions. Currently it includes images from Sentinel 1 -2- and -3.

The CALLISTO data repository engages the scientific community in the available opportunities for Copernicus data by generating annotated datasets that actually help their work. It is a collection of datasets from 4 themes: agriculture monitoring, water quality assessment, satellite journalism and land border change detection. It contains analysis ready remote sensing data with and without labels, in-situ and ground level datasets and geo-referenced labels.

### Volume of data

Copernicus is producing several terabytes of data every day. Moreover the analysis techniques going from simple statistics to the level of deep neural networks are numerous. The large volume of data combined with the potential of ML and DL has the potential to develop new applications rapidly and at large scale. However, the volume of the data and the expert knowledge needed to handle ML pipelines for EO data, transcends the capacity of many data scientists to extract meaningful information from them. One example is the analysis of SAR data which for many reasons could be a highly valuable source of information thanks to its ability to "see in the dark". The amplitude product of SAR is convenient for ML applications, such as object detection, as it 'feels' like traditional visible imagery. However it represents only a small portion of the entire SAR information content. Ignoring the phase content means that the complex physics of the backscattered SAR signal is entirely lost. However, trying to interpret the phase signal quickly becomes overwhelming to casual and expert data scientists. Hence, the EO community needs technologies from the ICT field.

DEEP CUBE combines mature and new ICT technologies, such as the Earth System Data Cube, the Semantic Cube, the Hopsworks platform for distributed Deep Learning, and a state-of-the-art visualization tool and integrates them to deliver an open and interoperable platform that can be deployed in several cloud infrastructures and High-Performance Computing, including the cloud-based platforms providing centralized access to Copernicus data (DIAS). According to expert knowledge it is important to further develop and centralize information, data, workflow pipelines and visualization tools on platforms in order to stay competitive with large cooperation with abundant resources such as Google, OpenAI and DeepMind. Such platforms should be user friendly and tailored to bring together both casual scientists and experts from the field of AI and EO.

### Explainability and causality

Artificial intelligence methodologies such as neural networks do not give any direct explanations why a given prediction or outcome was achieved by the network (Sarker *et al.*, 2017). This is due to the complicated transformations of input data and the algorithms themselves (Sarker *et al.*, 2020). The need to explain these outputs is called explainable artificial intelligence (XAI). XAI has many faces which include casual explainability which seeks to answer why a particular input gave the output in question and expert approach which looks how the AI model produced the output. According to expert knowledge, to obtain an element of explainability hence understanding of complex AI algorithms, traditional technologies such as knowledge graphs are used. In earth observation studies, in data analysis process, KGs encode human knowledge in machine-readable formats, which can be applied to aid data management and analysis (Ma, 2022). Moreover, by applying AI techniques on data selected through knowledge graphs, a keen understanding of the discrimination process of the knowledge graph can provide an insight on how an AI model achieved a given output (Ma, 2022).

Causality refers to a process, state, or a cause, which contributes to the production of another event, process or state, an effect, where the cause is partly responsible for the effect, and the effect is partly dependent on the cause. Causality is an important concept in studying the dynamic surface of our living planet however its integration in artificial intelligence methodologies used to study earth observation is still young. In a study by Otgonbaatar *et al.* (2022), the use of a causal directed acyclic graph could provide valuable insight on the cause-effect relation between cloud coverage and the agriculture land change as opposed to water quantity. However, this study used a simple causal structure where the causal relations are known. Where causal relationships are unknown and more data is used, discovering casual relations can be a problem. Therein lies one of the challenges affecting the integration of causality in AI models for earth observation.

These same factors controlling the potential use of AI, identified by stakeholders and experts, were also indicated in the literature review in section 4.2.

### 5.2.4 Technological Readiness Level

Technological Readiness Levels (TRL, originally developed by NASA) are used to determine the maturity of a technology. It offers a consistent way to evaluate the maturity of different available technologies and compare them amongst each other. Classic TRL's consist of 9 levels, level 1 providing the basic principles whilst level 9 considers the actual deployment of a technology/system (most mature level). Lavin *et al.* (2021) introduces a TRL framework, streamlined towards ML workflows. The Machine Learning Technological Readiness Levels (MLTRL) used to go from basic research on ML methods towards productization and deployment, by Lavin *et al.* (20121), are presented in Figure 10.

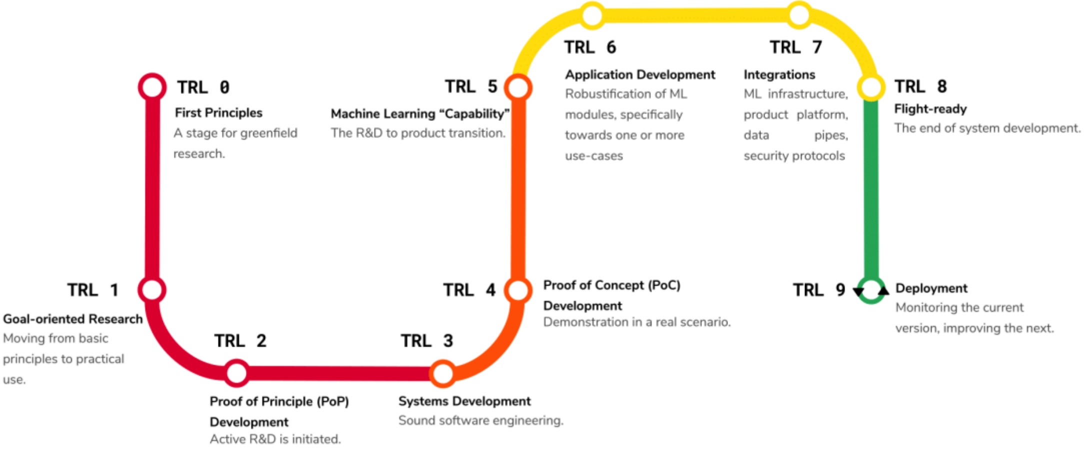

TRL 0
First Principles
A stage for greenfield research.
TRL 1
Goal-oriented Research
Moving from basic principles to practical use.
TRL 2
Proof of Principle (PoP) Development
Active R&D is initiated.
TRL 3
Systems Development
Sound software engineering.
TRL 4
Proof of Concept (PoC) Development
Demonstration in a real scenario.
TRL 5
Machine Learning "Capability"
The R&D to product transition.
TRL 6
Application Development
Robustification of ML modules, specifically towards one or more use-cases
TRL 7
Integrations
ML infrastructure, product platform, data pipes, security protocols
TRL 8
Flight-ready
The end of system development.
TRL 9
Deployment
Monitoring the current version, improving the next.

*Figure 10: Technological Readiness Levels streamlined towards ML workflows (Lavin et al., 2021).*

An issue perceived by the expert community (based on the consultation of stakeholders and literature review) are the existing limitations of the robustness of the AI formalisms currently used for a wide variety of EO applications. Moreover it seems that the EO community keeps turning to the same AI formalisms and that a large portion of the AI spectrum does not reach them. Many studies don't surpass the level of Proof of Concept (TRL 4) because the AI and EO community lack each other's experience and knowledge. The processing of SAR data is an example of this. SAR data can't easily be used for machine learning due to the intensive pre-processing and the required domain knowledge.

Furthermore, both communities are still in need of more and better Big Data management platforms with sustainable cloud solutions. This is illustrated by the fact that, also European scientists still turn to Google and Amazon platforms. It should be explored how DIAS platforms can further address these demands.

To these ends ICECUBE, an initiative by ICEEYE, is a first foray into open source data abstraction and tooling for SAR data handling and datacube creation. ICECUBE has rapidly evolved from a PoP (TRL 2) in 2021 to applying the library for many case studies and recently came out with a toolkit to build your own ICECUB. It's also working on a docker for ESA's SNAP software.

Furthermore, natural language processing is finding its way to aid to this need. A consortium with among others the university of Athens is working on a toolkit that processes the requirements of a user in natural language an delivers half-processed images that meet their needs. This project is currently still at its infancy.

The Sen2Like processor for a VC of Landsat and Sentinel data, is currently entering its pre-operational phase (TRL6) and will be offered to the EO community as an "on demand" processor. Validation tests on its accuracy have demonstrated that the quality information should be included as part of the delivered products. Improved quality information facilitates the use of multi-temporal data.

The commonly used AC algorithms that were described earlier in present report (Acolite, C2RCC, iCOR, l2gen, Polymer and Sen2Cor) are all considered mature but are still under active development. For open water common AC processors are adequate but for inland and coastal waters improvement is necessary.

### 5.2.5 Strategy & approach

Implementing ML algorithms in EO applications and data processing is one way to lead to higher quality applications or end user products. However, by tackling some of the bottlenecks (as identified in 5.2.3) of the EO data processing pipeline by using AI techniques, quality of data processing pipeline can be enhanced resulting in a cascading effect towards applications, final products and modelling results.

**Example 1**: when more labeled datasets would be available, better results for image fusion etc. can be obtained leading to higher resolutions and therefore higher quality and more accurate data products.

Recommendation: more open-source benchmark datasets

**Example 2**: by streamlining existing platforms and initiatives with regard to data download, ML pipelines and supporting documentation the use of AI for EO purposes can be optimized.

Recommendation: focusing on the need of integrating current platforms/projects

**Example 3**: by connecting the EO and AI community and assuring a better integration of the 2 domains difficulties with e.g. SAR data processing can be (partially) solved.

**Example 4:** Reducing the black box factor (increased explainability) could increase the user uptake of ML algorithms by EO experts.

On top of enhancing current important EO algorithms, support with the development of an entire data processing pipeline given the numerous data sources, -techniques and visualization options should be considered.

The above findings (previous sections) stress the importance of a holistic approach, instead of trying to find one-size-fits-all solutions.

# 6 Conclusions & recommendations

Based on literature and stakeholder and expert consultations the following conclusions are drawn:

- The **number of projects and initiatives** focusing on the need of the integration of EO data processing, -assimilation and application building and artificial intelligence techniques **is significant and rising**.

- The experts of the EO community acknowledge the **high potential of AI for EO**. **However,** at this moment each expert (or expert group) tends to have its **own "go to practices"** and they often have difficulties in explaining the outcomes of the AI based model. A lot of techniques and strategies are tested but **TRL's tend to remain low** in general (exceptions of high TRL ML based products exist).
  - A better integration of the EO and AI community would be beneficiary to solve these issues.

- **Bottlenecks** limiting the use of AI for optimal exploitation of EO data **are the lack of labeled datasets, the volume of data and the explainability/causality of events.**

- **Deep learning techniques** have proven to immensely push the mining of aquatic remote sensed data forward. However, deep learning algorithms are challenging, especially to scientists that are unexperienced in this field. Data platforms tailored to developing ML pipelines for EO data should consider adding example code or even clearn and (to the possible extent) easy to understand video tutorials elaborating on commonly encountered issues when dealing with machine/deep learning algorithms.

- **Standard validation protocols** must be established in order to clean in situ datasets. Also these could be included in the current DIAS platform in the form of notebooks/tutorials/SOPs,..
- **Standard satellite data processing guidelines** are oftentimes missing for experts from various backgrounds. This would be a big leap forward in order to train ML algorithms with higher accuracy.
- **Validated in situ data or dataset builders** to create ready-to-use training datasets (i.e., in situ and satellite data pairing) might help AI researchers which might not have EO expertise to create more advanced AI/ML methods.
- The **lack of streamlined platforms and initiatives** with regard to data download, ML pipelines and supporting documentation for the use of AI for EO purposes can be optimized. The platforms that exist today are not equipped to guide researchers from all walks of life, with all levels of experience to develop a plan of action. For researchers that are new to e.g. de DIAS platforms it is a knot to untangle that may push them to other well-established platforms.
- It's often too **complicated to find the (expert) relevant information** and training material to make optimal use of the existing platforms. Although this information is available, it's scattered and not tailored to the different skill levels of the experts. One could think for example about a bot guiding new users through the DIAS platforms, adding links to information that is tailored to different levels of expertise, adding example code for example in the form of notebooks that are well documented,...
- **Focus on a holistic approaches** could be beneficial instead of trying to find one-size-fits-all solutions. One could think for example about developing SOPs or dashboards that help researchers develop their research plan. E.g. in the case of AC, we think it is necessary to further improve on the techniques discussed above and push science forward. However, one could argue that helping a researcher choose the optimal, already existing algorithm for their research question at hand would in the meantime be a significant step forward. The collection and pre-processing (i.e., geometric and atmospheric correction) of satellite data is a time-consuming

operation. The pre-processing steps and algorithms applied might differ between various implementations and affect the performance of the implemented AI method.

# 7 References

Akhand, K., Nizamuddin, M., Roytman, L., Kogan, F., (2016) "Using remote sensing satellite data and artificial neural network for prediction of potato yield in Bangladesh". Proc. SPIE, Remote Sensing and Modeling of Ecosystems for Sustainability, 9975, 997508.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., & Notarnicola, C. (2015) "Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data". Remote Sensing, 7(12), 16398-16421.

Andries, A., Morse, S., Murphy, R., Lynch, J., Woolliams, E., Fonweban, J. (2019) "Translation of Earth observation data into sustainable development indicators: An analytical framework". Sustainable Development, 27(3), 366-376.

Bose, P., Kasabov, N. K., Bruzzone, L., Hartono, R. N. (2016) "Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series". In IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 11, pp. 6563-6573. https://doi: 10.1109/TGRS.2016.2586602.

Brockmann, C., Doerffer, R., Peters, M., Kerstin, S., Embacher, S., Ruescas, A. (2016) "Evolution of the C2RCC Neural Network for Sentinel 2 and 3 for the Retrieval of Ocean Colour Products in Normal and Extreme Optically Complex Waters". Living Planet Symposium, Proceedings of the conference held 9-13 May 2016 in Prague, Czech Republic. Edited by L. Ouwehand. ESA-SP Volume 740, ISBN: 978-92-9221-305-3, p.54.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018) "A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach". Remote sensing of environment, 210, 35-47.

Cheng, G., Han, J. (2016). "A survey on object detection in optical remote sensing images". ISPRS Journal of Photogrammetry and Remote sensing, 117, 11-28.

Costache, R., Pham, Q.B., Corodescu-Ros, E., Cîmpianu, C., Hong, H., Thi Thuy Linh, N., Ming Fai, C., Ahmed, A., Vojtek, M., Pandhiani, S.M., Minea, G., Ciobotaru, N., Popa, M.C., Diaconu, D.C, Pham B.T. (2020) "Using GIS, Remote Sensing, and Machine Learning to Highlight the Correlation between the Land-Use/Land-Cover Changes and Flash-Flood Potential". Remote Sensing, 12, 1422; https://doi:10.3390/rs12091422.

Deng Z., Sun H., Zhou S., Zhao J., Lin Lei, Huanxin Zou. (2018) "Multi-scale object detection in remote sensing imagery with convolutional neural networks". ISPRS Journal of Photogrammetry and Remote Sensing, Volume 145, Part A, pages 3-22, https://doi.org/10.1016/j.isprsjprs.2018.04.003.

Ding, P., Zhang, Y., Deng, W. J., Jia, P., & Kuijper, A. (2018) "A light and faster regional convolutional neural network for object detection in optical remote sensing images". ISPRS journal of photogrammetry and remote sensing, 141, 208-218.

Eljasik-Swoboda, T., Rathgeber, C., & Hasenauer, R. (2019) "Assessing Technology Readiness for Artificial Intelligence and Machine Learning based Innovations". International Conference on Data Technologies and Applications.

Ennouri, K., Smaoui, S., Gharbi, Y., Cheffi, M., Braiek, O., Ennouri, M., Ali Triki, M. (2021) "Usage of Artificial Intelligence and Remote Sensing as Efficient Devices to Increase Agricultural System Yields". Journal of Food Quality, vol. 2021, 17 pages, Article ID 6242288. https://doi.org/10.1155/2021/6242288.

Ferreira, B., Iten, M., Silva, R.G. (2020) "Monitoring sustainable development by means of earth observation data and machine learning: a review". Environ Sci Eur, 32, 120. https://doi.org/10.1186/s12302-020-00397-4.

Fuentes, S., Tongson, E.J., De Bei, R., Gonzalez Viejo, C., Ristic, R., Tyerman, S., Wilkinson, K. (2019) "Non-Invasive Tools to Detect Smoke Contamination in Grapevine Canopies, Berries and Wine: A Remote Sensing and Machine Learning Modeling Approach". Sensors, 19, 3335; https://doi:10.3390/s19153335.

GEO (2017) "Earth Observations in support of the 2030 Agenda for Sustainable Development". Retrieved from https://www.earthobservations.org/documents/publications/201703_geo_eo_for_2030_agend a.pdf

Giles, Anna & Davies, James & Ren, Keven & Kelaher, Brendan. (2021) "A deep learning algorithm to detect and classify sun glint from high-resolution aerial imagery over shallow marine environments". ISPRS Journal of Photogrammetry and Remote Sensing. 181. 20-26. 10.1016/j.isprsjprs.2021.09.004.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B. (2017) "SoilGrids250m: global gridded soil information based on machine learning". PLoS ONE 12, e0169748. https://doi.org/10.1371/journal.pone.0169748

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., Zhu, X. (2018) "Identifying corresponding patches in SAR and optical images with a Pseudo-Siamese CNN". IEEE Geosci. Remote Sens. Lett., 15 (5), pp. 784-788.

Jentzsch, S., Hochgeschwender, N. (2021) "A qualitative study of Machine Learning practices and engineering challenges in Earth Observation" IT - Information Technology, vol. 63, no. 4, pp. 235-247. https://doi.org/10.1515/itit-2020-0045.

Jin, X., Li, Z., Feng, H., Renc, Z., Li, S. (2019) "Deep neural network algorithm for estimating maize biomass based on simulated Sentinel 2A vegetation indices and leaf area index". The Crop Journal, 8, 87–97.

Knipper, K. R., Kustas, W. P., Anderson, M. C., Alfieri, J. G., Prueger, J. H., Hain, C. R. (2019) "Evapotranspiration estimates derived using thermal-based satellite remote sensing and data fusion for irrigation management in California vineyards". Irrig. Sci., 37(3), 431-449.

Lary, D. J., Zewdie, G. K., Liu, X., Wu, D., Levetin, E., Allee, R. J., Malakar, N., Walker, A., Mussa, H., Mannino, A., Aurin, D. (2018) "Machine learning applications for earth observation". Earth observation open science and innovation, 165.

Lavin, A., Gilligan-Lee, C.M., Visnjic, A. (2022) "Technology readiness levels for machine learning systems". Nat. Commun. 13, 6039. https://doi.org/10.1038/s41467-022-33128-9.

Li, J., He, Z., Plaza, J., Li, S., Chen, J., Wu, H., Liu, Y. (2017) "Social media: New perspectives to improve remote sensing for emergency response". Proceedings of the IEEE, 105(10), 1900-1912.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). "Deep learning in remote sensing applications: A meta-analysis and review". ISPRS journal of photogrammetry and remote sensing, 152, 166-177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.

Merkle, N., Auer, S., Müller, R., Reinartz, P. (2018) "Exploring the potential of conditional adversarial networks for optical and SAR image matching". IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 11 (6), pp. 1811-1820.

Meyer, Franz. "Spaceborne Synthetic Aperture Radar – Principles, Data Access, and Basic Processing Techniques." SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. Eds. Flores, A., Herndon, K., Thapa, R., Cherrington, E. NASA

Nascimento, E., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., Conte, T. (2019) "Understanding development process of machine learning systems: Challenges and solutions". ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), pp. 1-6.

Otgonbaatar, Soronzonbold & Datcu, Mihai & Demir, Begüm. (2022) "Causality for Remote Sensing: An Exploratory Study". 10.1109/IGARSS46834.2022.9883060.

Pahlevan, N., Greb, S., Dekker, A.G. (2022) "Earth Observation in Support of SDG 6.3.2/6.6.1". Geophysical Monograph Series, chapter 4. https://doi.org/10.1002/9781119536789.ch4.

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., Alikas, K., Kangro, K., Gurlin, D., Hà, N., Matsushita, B., Moses, W., Greb, S., Lehmann, M.K., Ondrusek, M., Oppelt, N., Stumpf, R. (2020). "Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and

Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach". Remote Sensing of Environment, 240, 111604.

Pahlevan N., Mangin A., Sundarabalan V. Balasubramanian, Smith B., Alikas K., Arai K., Barbosa C., Bélanger S., Binding C., Bresciani M., (2021) ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters, Remote Sensing of Environment, https://doi.org/10.1016/j.rse.2021.112366

Pan, Y.; Bélanger, S.; Huot, Y. (2022) "Evaluation of Atmospheric Correction Algorithms over Lakes for High-Resolution Multispectral Imagery: Implications of Adjacency Effect. Remote Sens" , 14, 2979., https://doi.org/10.3390/rs14132979

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. (2019) "Deep learning and process understanding for data-driven Earth system science". Nature, 566, 195-2004.

Rosser, J. F., Leibovici, D. G., Jackson, M. J. (2017) "Rapid flood inundation mapping using social media, remote sensing and topographic data". Nat. Hazards, 87(1), 103-120.

Saunier, S., Pflug, B., Lobos, IM., Franch, B., Louis, J., De Los Reyes, R., Debaecker, V., Cadau, EG., Boccia, V., Gascon, F., Kocaman, S., Sen2Like. (2022) "Paving the Way towards Harmonization and Fusion of Optical Data". Remote Sensing, 14(16):3855. https://doi.org/10.3390/rs14163855.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., Adams, C. (2020) "Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing". Earth-Science Reviews, 205, 103187.

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-MartÃnez, A., Izquierdo-Verdiguier, E., MuÃ±oz-MarÃ, J., Mosavi, A., Camps-Valls, G. (2020). "Machine Learning Information Fusion in Earth Observation: A Comprehensive Review of Methods, Applications and Data Sources". Information Fusion, S1566253520303171. https://doi:10.1016/j.inffus.2020.07.004.

Sarkar A., Yang Y., Vihinen M. (2020) "Variation benchmark datasets: update, criteria, quality and applications", Volume 2020, https://doi.org/10.1093/database/baz117

Schiller, H., Doerffer, R. (1999) "Neural network for emulation of an inverse model operational derivation of Case II water properties from MERIS data". International Journal of Remote Sensing, 20:9, 1735-1746, https://doi.org/10.1080/014311699212443.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M. (2014) "Machine learning: The high interest credit card of technical debt". Google, Inc.

Smith, B., Pahlevan, N., Schalles, J., Ruberg, S., Errera, R., Ma, R., Giardino, C., Bresciani, M., Barbosa, C., Moore, T., Fernandez, V., Alikas K., Kangro, K. (2021). "A chlorophyll-a algorithm for Landsat-8 based on mixture density networks". Frontiers in Remote Sensing, 1, 623678.

Sorkhabi, O. M., Shadmanfar, B., Kiani, E. (2022). "Monitoring of dam reservoir storage with multiple satellite sensors and artificial intelligence. Results in Engineering, 16, 100542".

Sun, A. Y., & Scanlon, B. R. (2019) "How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions". Environmental Research Letters, 14(7), 073001.

Tan, J., NourEldeen, N., Mao, K., Shi, J., Li, Z., Xu, T., Yuan, Z. (2019) "Deep Learning Convolutional Neural Network for the Retrieval of Land Surface Temperature from AMSR2 Data in China". Sensors, 19, 2987. https://doi.org/10.3390/s19132987.

TowardsDataScience, consulted last on 16/11/2022. https://medium.com/towards-data-science

Topouzelis K., Papageorgiou D., Giuseppe Suaria, Stefano Aliani (2021) "Floating marine litter detection algorithms and techniques using optical remote sensing data: A review" Marine Pollution Bulletin, Volume 170, https://doi.org/10.1016/j.marpolbul.2021.112675.

Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J., Moreno, J. (2012) “Retrieval of vegetation biophysical parameters using Gaussian process techniques”. IEEE Trans. Geosci. Remote Sens., 50, 1832–1843.

Volke, M. I., Abarca-Del-Rio, R. (2020) “Comparison of machine learning classification algorithms for land cover change in a coastal area affected by the 2010 Earthquake and Tsunami in Chile”. Nat. Hazards Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/nhess-2020-41.

Wang, H., Qi, J., Lei, Y, Wu, J., Li, B., Jia, Y. A. (2021) “Refined Method of High-Resolution Remote Sensing Change Detection Based on Machine Learning for Newly Constructed Building Areas”. Remote Sens., 13, 1507. https://doi.org/10.3390/rs13081507.

Wang, L., Dong, Q., Yang, L. (2019) “Crop classification based on a novel feature filtering and enhancement method. Remote Sens, 11, 455. https://doi.org/10.3390/rs11040455.

Warren, M., Simis, S., Martinez-Vicente, V., Poser, K., Bresciani, M., Alikas, K., Spyrakos, E., Giardino, C., Ansper, A. (2019) “Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral Imager over coastal and inland waters”. Remote Sensing of Environment. 225. 267-289. 10.1016/j.rse.2019.03.018.

Water-ForCE. (2022b) “D1.6: Report on links and gaps between satellite EO and water related SDGs and climate indicators”.

Water-ForCE. (2022a) “D1.4: Report with end-user needs and requirements”.

Yang, Y., Yang, J., Xu, C., Xu, C., Song, C. (2019) “Local-scale landslide susceptibility mapping using the BGeoSVC model” Landslides, 16(7), 1301-1312.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L. (2020). “Deep learning in environmental remote sensing: Achievements and challenges”. Remote Sensing of Environment, 241, 111716. https://doi.org/10.1016/j.rse.2020.111716.

Zerrouki, N., Harrou, F., Sun, Y., & Hocini, L. (2019) “A machine learning-based approach for land cover change detection using remote sensing and radiometric measurements”. IEEE Sensors Journal, 19(14), 5843-5850.

Zhao, Xin, Hui Li, Ping Wang, and Linhai Jing. 2021. ”An Image Registration Method Using Deep Residual Network Features for Multisource High-Resolution Remote Sensing Images” *Remote Sensing* 13, no. 17: 3425. https://doi.org/10.3390/rs13173425

Zhang, B., Gu, J., Chen, C., Han, J., Su, X., Cao, X., Liu, J. (2018) “One-two-one networks for compression artifacts reduction in remote sensing”. ISPRS journal of Photogrammetry and Remote sensing, 145, 184-196.

# 8 Annexes

Annex 1: Overview of 22 consulted companies for EO-AI community feedback.

Annex 2: Companies consulted by means of a survey (+survey questions below). Responsive companies are indicated in green.

Annex 3: Results of the Mentimeter poll during the Water-ForCE webinar on the 26th of October, 2022.

*Annex 1: Overview of 22 consulted companies for EO-AI community feedback. Responsive companies are indicated in green.*

| Name company | Description core business | Link |
|---|---|---|
| Amigo | Climate data analytics. | https://amigoclimate.com/#NUA |
| Artificial Intelligence Data Analysis (Horizon2020) | AIDA developed a new open source software called AIDApy written in Python (a free language) and capable of collecting, combining and correlating data from different space missions. | Home \| AIDA (aida-space.eu) |
| ARTificial Intelligence for Seasonal forecast of Temperature extremes | Project seeks to improve insights into climate predictability at the seasonal timescale, aiming to increase the performance of existing prediction systems. | https://cordis.europa.eu/project/id/101033654 |
| CLImate INTelligence: Extreme events detection, attribution and adaptation design using machine learning (Horizon2020) | Involved in the development of an Artificial Intelligence framework composed of Machine Learning techniques and algorithms to process big climate datasets for improving Climate Science in the detection, causation, and attribution of Extreme Events (EEs), namely tropical cyclones, heatwaves and warm nights, droughts, and floods. | https://climateintelligence.eu/ |
| CloudFerro | Provides cloud computing services with a focus on big data sets. Contributed to WEkEO and operates CREADIAS. | https://cloudferro.com/en/ |
| DigiFarm AS / ALTYN Sarl / Farmen Gard | Sen4Weeds project will develop a solution for automated large-scale detection and mapping of weeds in agricultural fields using remote sensing and artificial intelligence. | https://ai4copernicus-project.eu/sen4weeds-automatic-detection-and-mapping-of-in-field-weeds/ |
| Earth Science Data Systems | Program which promotes the use AI and recognizes its potential to significantly advance existing data systems capabilities, improve operations, and maximize the use of NASA Earth observing data. | https://www.earthdata.nasa.gov/esds/ai-ml |
| ECMWF | Earth observation, linked to AI4Copernicus | https://www.ecmwf.int/en/computing |

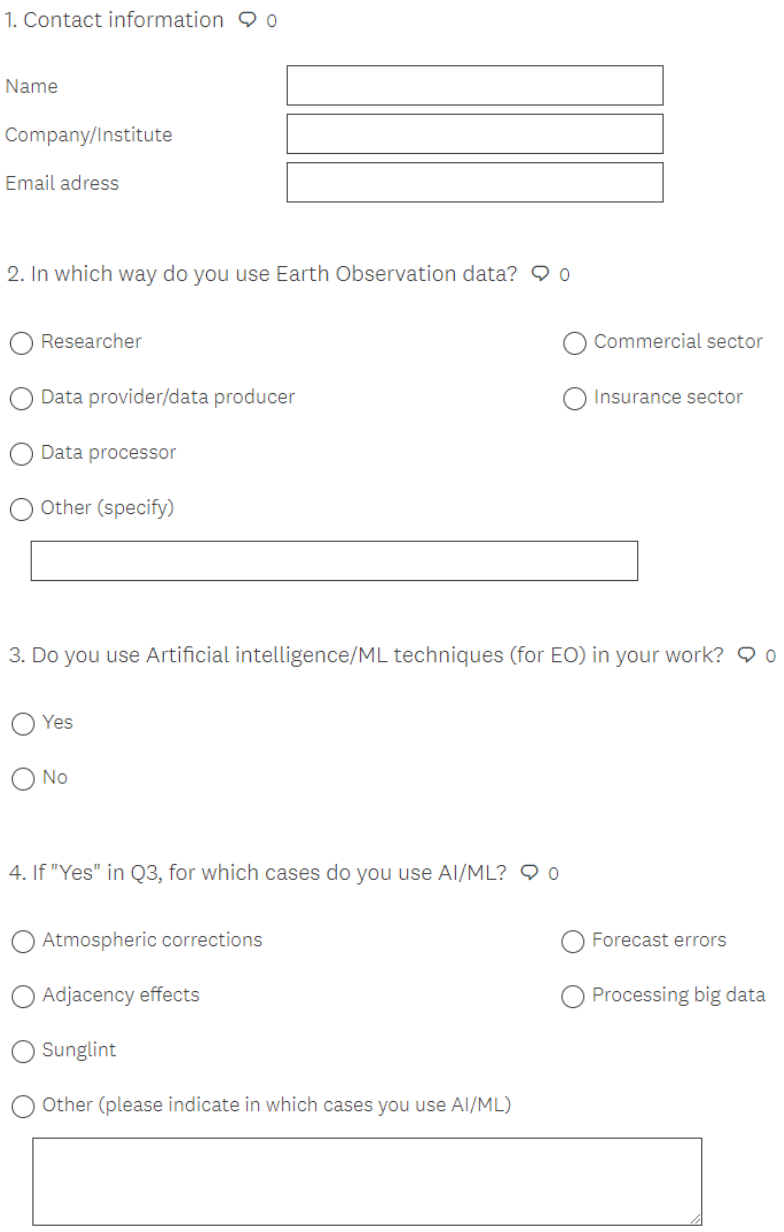
| | | |
|---|---|---|
| EO Science for society | EO Science for society of ESA | https://eo4society.esa.int/ |
| European union satellite centre | Geospatial intelligence - exploitation and analysis of imagery and geospatial information to describe, assess, and visually depict physical features and geographically referenced activities on Earth. | https://www.satcen.europa.eu/services/research_technology_development_and_innovation |
| GEO.INFORMED | Develops deep learning workflows that can transform Copernicus Sentinel data into the information that is needed by environmental policy agencies. | GEO.INFORMED – Remote sensing & Deep learning for environmental policy (geo-informed.be) |
| Institute of Informatics & Telecommunications: Computational Intelligence Laboratory | AI for geoapplications. | https://www.iit.demokritos.gr/labs/cil/ |
| Institute of Informatics & Telecommunications: Software & Knowledge Engineering Lab | Reinforcing the AI4EU platform by advancing earth observation intelligence, innovation and adoption. | https://www.skel.ai/skel-projects/#! |
| JOANNEUM RESEARCH - DIGITAL | Institute for Information and Communication Technologies. Develops applied high tech solutions for the following markets: Mobility, Space, Industry, Security & Defence, Energy & Environment, application-oriented research partner. | https://www.joanneum.at/en/life/research-areas/weather-risk-analysis-and-management |
| Management of Data Information and knowledge group, Department of informatics and telecommunications National and | Involved in artificial intelliegence for earth observation, have developed programes used in the copernicus program. | AI Team (uoa.gr) |

| | | |
|---|---|---|
| Kapodistrian University of Athens | | |
| Planetary computer | Catalog of global environmental data with intuitive APIs | Home \| Planetary Computer (microsoft.com) |
| Remote sensing laboratory, university of Trento | Reinforcing the AI4EU Platform by advancing earth observation intelligence, innovation and adoption of advanced machine learning techniques | https://rslab.disi.unitn.it/contact/ |
| SCAVIHO – Scalable Vegetation Index and Harvesting Forecaster | Focuses on NVDI and precision agriculture, linked to AI4Copernicus | https://ai4copernicus-project.eu/scaviho-scalable-vegetation-index-and-harvesting-forecaster/ |
| SISTEMA GmbH and cmc-consulting | The "Super Resolution for Climate Crisis Context – SR4C3" AI4Copernicus project aims at bringing innovation to the climate crisis sector by enhancing the remote sensing based technological tools through the application of Artificial Intelligence | http://www.sistema.at/wp/ |
| Solaïs and Transvalor | Through the project satelite images prediction with deep learning, this project will Develop satellite images forecasting techniques using advanced Deep Learning so as is to improve short term solar irradiation forecasts in the 15 minutes to 6 hours range | https://ai4copernicus-project.eu/sen4weeds-automatic-detection-and-mapping-of-in-field-weeds/ |
| TNO | TNO is a leading partner in the Dutch AI Coalition. | https://www.tno.nl/en/about-tno/our-people/alexander-eijk/ |
| Trasys | Has worked with ESA and VITO on earth observation projects | http://www.trasysinternational.com/projects/ |
| VTT | VTT is a research, development and innovation partner for organizations in the space industry. | https://www.vttresearch.com/en/industries/space-industry |

*Annex 2: Companies consulted by means of a survey (+survey questions below). Responsive companies are indicated in green.*

| Name company | Description core business | Link |
|---|---|---|
| Artys S,r,l | EO4NOWCAST – Earth Observation for Severe Weather Hazard Nowcasting ambition is to realise and demonstrate an operational and replicable approach to assess severe weather events and related hazards in the short term (nowcasting) built upon the synergy between EO and rainfall monitoring products. | https://www.artys.it/en/ |
| Impact observatory | Impact Observatory brings AI-powered algorithms and on-demand data to sustainability and environmental risk analysis for governments, industries, and markets. | Impact Observatory |
| Latitudo 40 | Urbalytics - The proposed project aims at demonstrating the possibility of applying machine/deep learning algorithms on Sentinel 2 images in order to estimate the Land Surface Temperature. | https://www.latitudo40.com/technology/ |
| STAM S.r.l. / Gter S.r.l. / La SIA S.p.A | 4th open call AI4Copernicus winner, Their winning project is AI-RON MAN – AI-based wildfiRe predictiON for the risk MANagement of TLC Infrastructures. | |
| 3D Executive Management System / List Geoinformatika / Profida | AI4 E2O.Green will develop an Intelligent next gen deep/green tech Platform powered by AI to enable Urban Green and Golf Space Management Companies to effectively manage irrigation, assets, operations and land fields with a powerful combination of satellite and drones imagery as well as the AR and computer vision connected models. | https://ai4copernicus-project.eu/4th-round-of-open-calls-meet-the-winners/ |

## Survey questions:



1. Contact information 🗨 0

Name

Company/Institute

Email adress

2. In which way do you use Earth Observation data? 🗨 0

○ Researcher

○ Data provider/data producer

○ Data processor

○ Other (specify)

○ Commercial sector

○ Insurance sector

3. Do you use Artificial intelligence/ML techniques (for EO) in your work? 🗨 0

○ Yes

○ No

4. If "Yes" in Q3, for which cases do you use AI/ML? 🗨 0

○ Atmospheric corrections

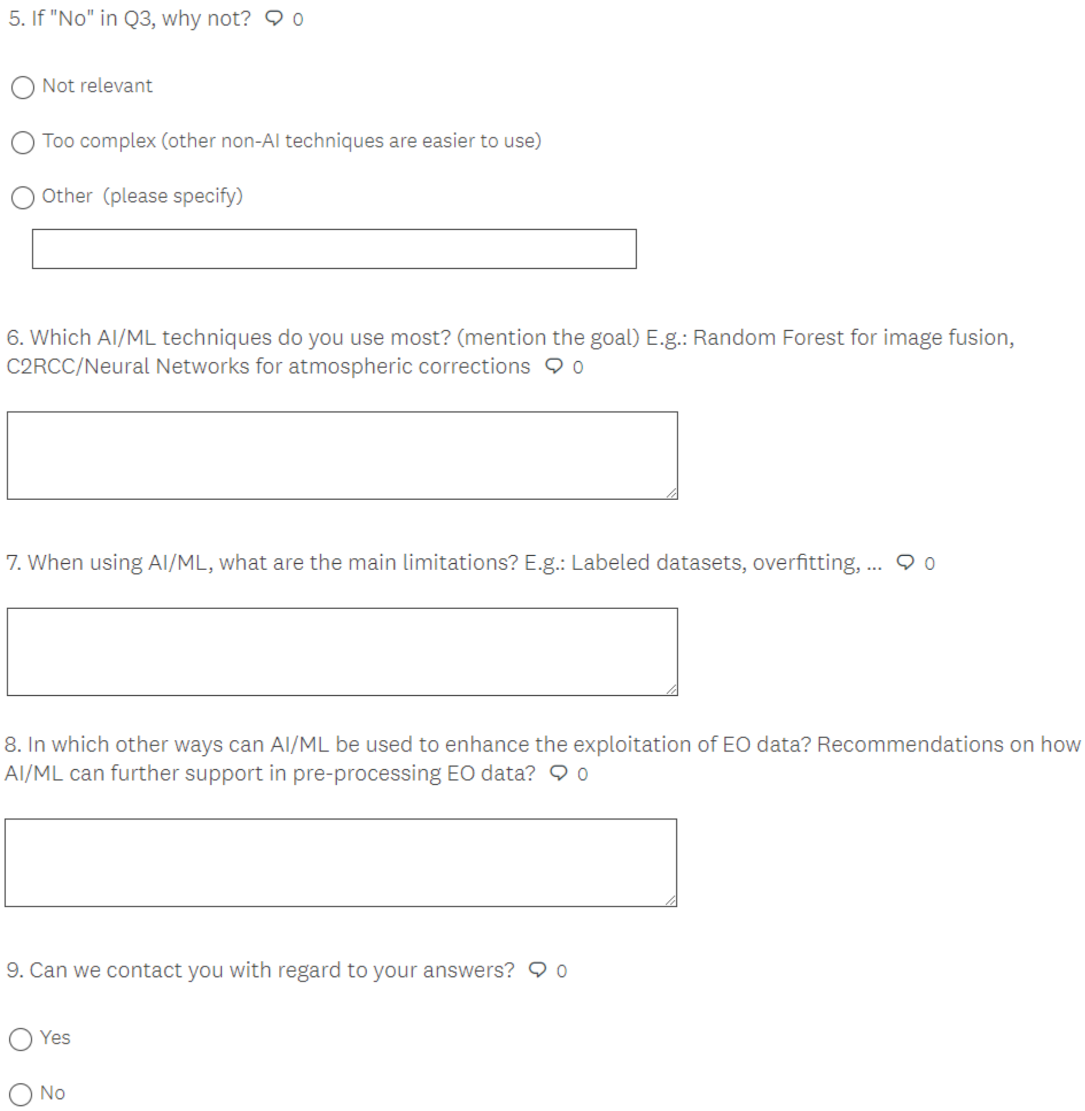
○ Adjacency effects

○ Sunglint

○ Other (please indicate in which cases you use AI/ML)

○ Forecast errors

○ Processing big data



5. If "No" in Q3, why not? 🗨 0

○ Not relevant

○ Too complex (other non-AI techniques are easier to use)

○ Other (please specify)

6. Which AI/ML techniques do you use most? (mention the goal) E.g.: Random Forest for image fusion, C2RCC/Neural Networks for atmospheric corrections 🗨 0

7. When using AI/ML, what are the main limitations? E.g.: Labeled datasets, overfitting, ... 🗨 0

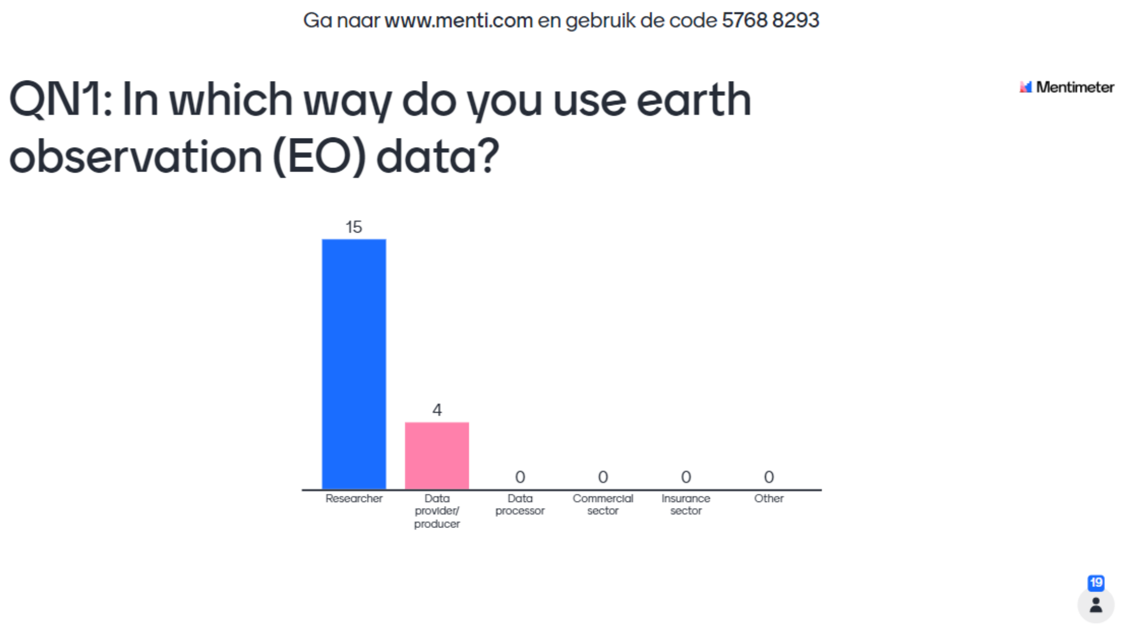
8. In which other ways can AI/ML be used to enhance the exploitation of EO data? Recommendations on how AI/ML can further support in pre-processing EO data? 🗨 0

9. Can we contact you with regard to your answers? 🗨 0

○ Yes

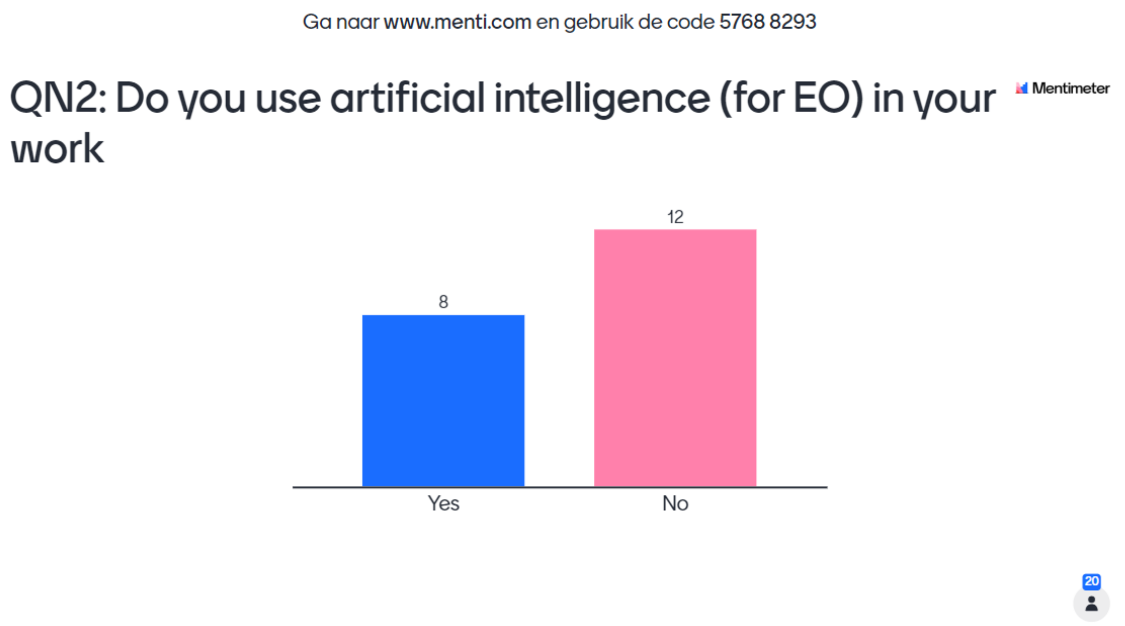
○ No

*Annex 3: Results of the Mentimeter poll during the Water-ForCE webinar on the 26th of October, 2022.*



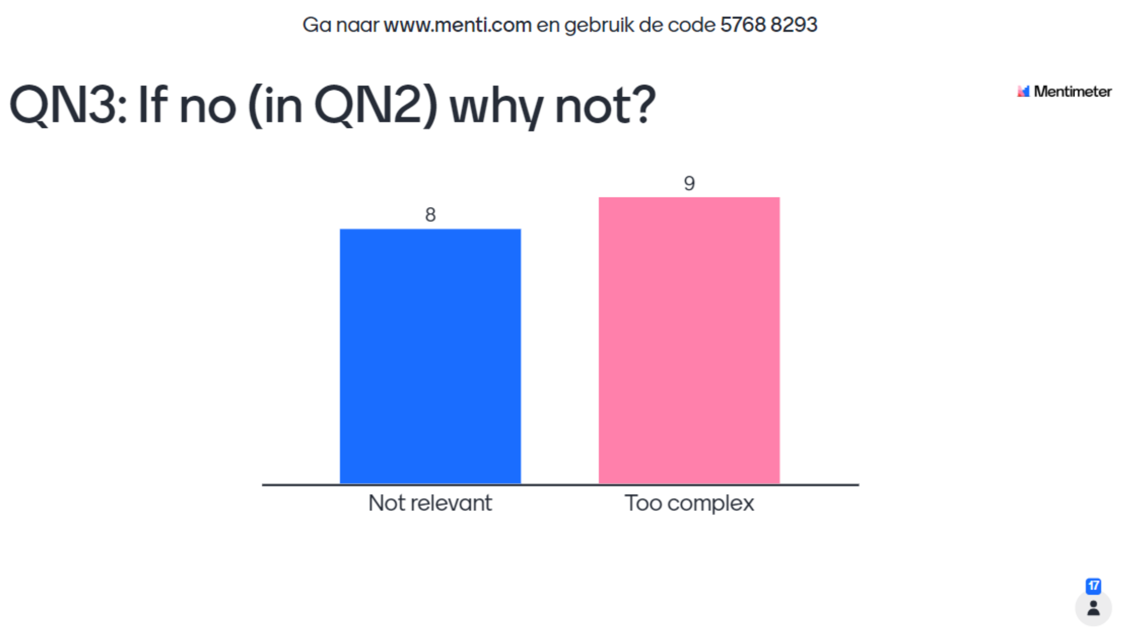
Ga naar www.menti.com en gebruik de code 5768 8293

QN1: In which way do you use earth observation (EO) data?



Ga naar www.menti.com en gebruik de code 5768 8293

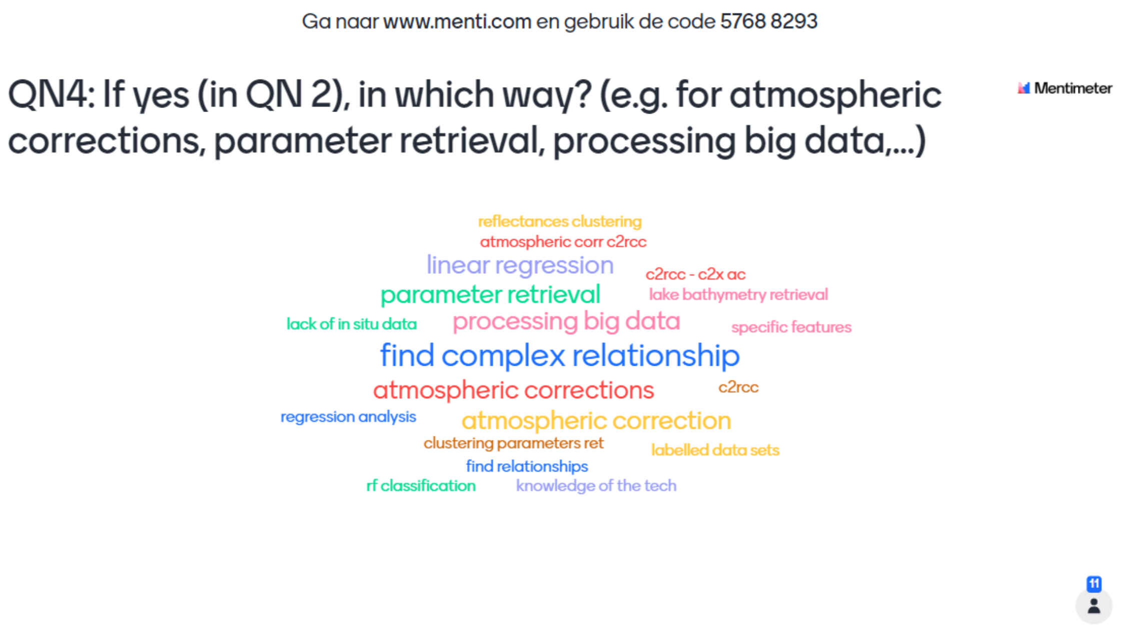
QN2: Do you use artificial intelligence (for EO) in your work



Ga naar www.menti.com en gebruik de code 5768 8293

## QN3: If no (in QN2) why not?

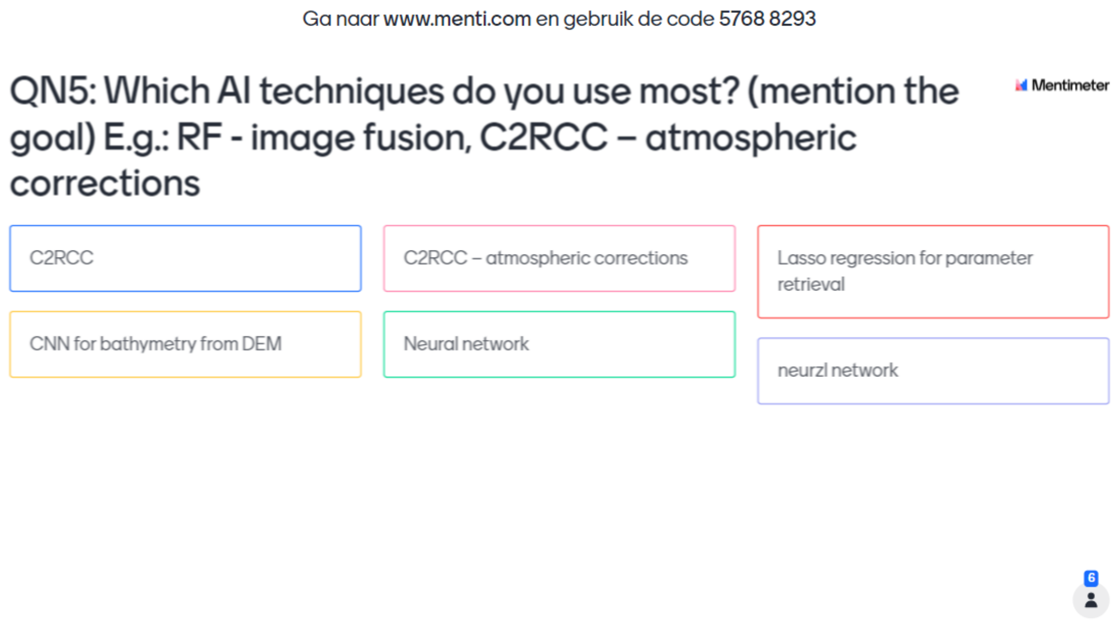
Mentimeter



Ga naar www.menti.com en gebruik de code 5768 8293

## QN4: If yes (in QN 2), in which way? (e.g. for atmospheric corrections, parameter retrieval, processing big data,...)

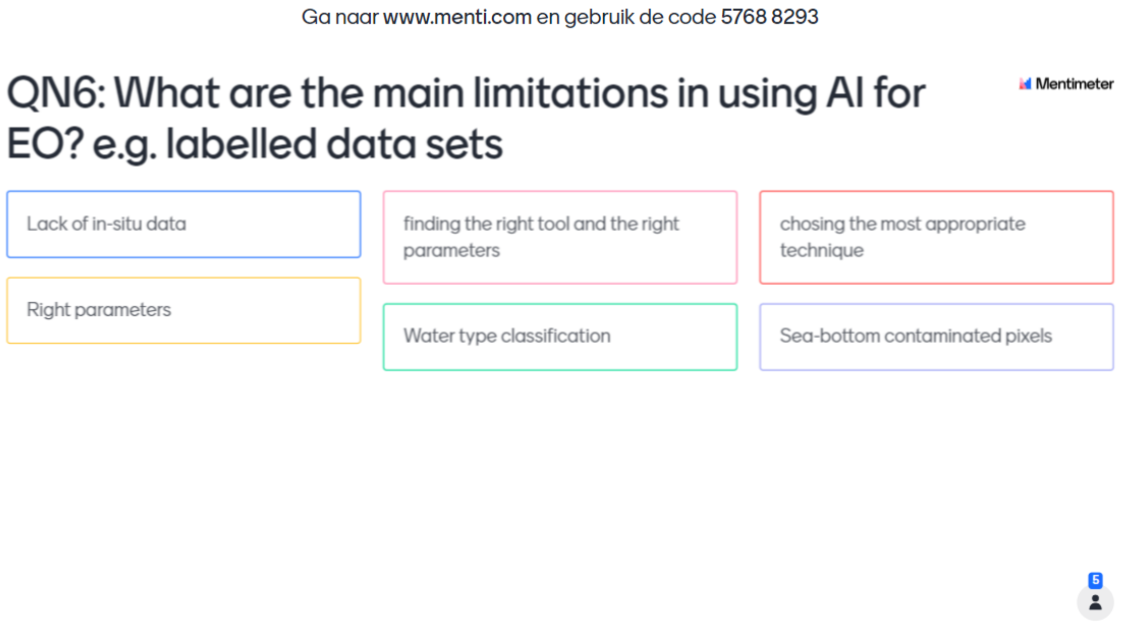
Mentimeter



Ga naar www.menti.com en gebruik de code 5768 8293

QN5: Which AI techniques do you use most? (mention the goal) E.g.: RF - image fusion, C2RCC – atmospheric corrections

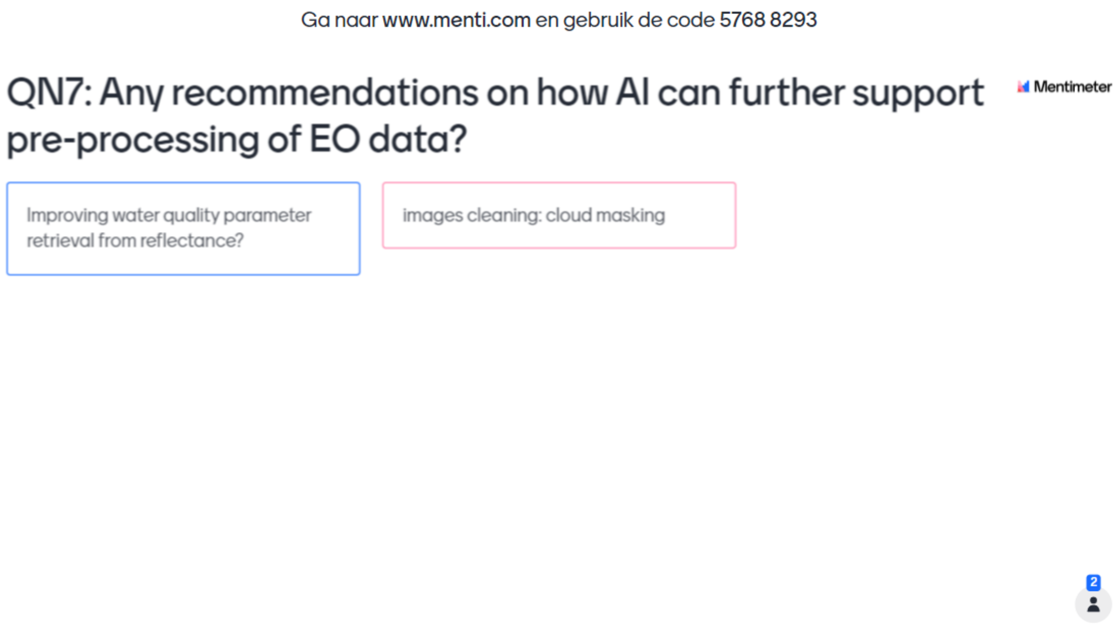
- C2RCC
- CNN for bathymetry from DEM
- C2RCC – atmospheric corrections
- Neural network
- Lasso regression for parameter retrieval
- neurzl network



Ga naar www.menti.com en gebruik de code 5768 8293

QN6: What are the main limitations in using AI for EO? e.g. labelled data sets

- Lack of in-situ data
- Right parameters
- finding the right tool and the right parameters
- Water type classification
- chosing the most appropriate technique
- Sea-bottom contaminated pixels



Ga naar www.menti.com en gebruik de code 5768 8293

QN7: Any recommendations on how AI can further support pre-processing of EO data?

Improving water quality parameter retrieval from reflectance?

images cleaning: cloud masking