EPN 2024 RI

EUROPLANET 2024 Research Infrastructure

H2020-INFRAIA-2019-1

Europlanet 2024 RI has received funding from the European Union's Horizon 2020 Research and Innovation Programme under
Grant agreement no: 871149

Deliverable D10.1
ML 1st year report

Due date of deliverable:	31/01/2021
Actual submission date:	30/01/2021
Nature:[1]	R
Dissemination level:[2]	PU
Work package:	WP10
Lead beneficiary:	IWF-OEAW
Contributing beneficiaries:	KNOW, UNIPASSAU, ACRI-ST, DLR, INAF, IAP-CAS, AOP
Document status:	Final

Start date of project: 01 February 2020. Duration: 48 months
Project Co-ordinator: Prof Nigel Mason, University of Kent

Executive Summary / Abstract:

This annual report summarises the work done in Work Package 10 ‘Machine Learning Solutions for Data Analysis and Exploitation in Planetary Sciences’ during the first year of Europlanet 2024 Research Infrastructure. The main aims of the work package are to foster wider use of machine learning technologies in data driven space research and to provide open source machine learning tools developed for specific science cases. Work Package 10 is organized around six tasks that target management and coordination of the activities, development of machine learning based data analysis algorithms, and dissemination of the tools as well as integration of the results into VESPA, GMAP and SPIDER. Despite the onset of the Covid-19 pandemic right at the beginning of the project, work on all of the six tasks has been progressing

1. **Nature:** R = Report, P = Prototype, D = Demonstrator, O = Other

2. **Dissemination level:**

PU	PP	RE	CO
Public	Restricted to other programme participants (including the Commission Service)	Restricted to a group specified by the consortium (including the Commission Services)	Confidential, only for members of the consortium (excluding the Commission Services)

and three science cases will be finished in February and March 2021. A so-called Machine Learning Portal has been installed and serves as an access point for our activities. Descriptions of the science cases, general information about machine learning activities within Europlanet, links to presentations and announcements of upcoming events are already available on the portal. More content will be added in the upcoming months. Further, a public GitHub organisation was set up, where codes and scripts are available for the scientific community. First results of the science cases were presented at the Europlanet Science Congress as well as the European Space Weather week.

TABLE OF CONTENTS

Table of Abbreviations	
BS	Bow Shock
CIR	Corotating Interaction Region
CME	Coronal Mass Ejection
CNN	Convolutional Neural Network
D	Deliverable
DMP	Data Management Plan
DTM	Digital Terrain Model
EGI	European Grid Infrastructure
EGU	European Geophysical Union
EOSC	European Open Science Cloud
EPN2024-RI	Europlanet 2024 Research Infrastructure
EPSC	EuroplanetScience Congress
ESWS	European Space Weather Symposium
GAN	Generative Adversarial Network
GMAP	Geologic MApping of Planetary bodies
ICME	Interplanetary Coronal Mass Ejection
IMF	Interplanetary Magnetic Field
JRA	Joint Research Activity
ML	Machine Learning
MP	Magnetopause
MS	Milestone
SPIDER	Sun Planet Interactions Digital Environment on Request
TRL	Technology Readiness Level
VA	Virtual Access
VESPA	Virtual European Solar and Planetary Access
WP	Work Package

1. Overview of the Work Package and Explanation of Work

1.1 Objectives and Description of Work

The objectives and description of work for Work Package (WP) 10 ‘JRA4 ML - Machine Learning Solutions for Data Analysis and Exploitation in Planetary Sciences’ are as follows, quoted from the proposal:

> JRA4 will develop Machine Learning (ML) powered data analysis and exploitation tools optimised for planetary science and integrate expert knowledge on ML into the planetary community. All tools will also be linked via the VA services of VESPA, GMAP and SPIDER (where appropriate).
>
> The main objectives are:
>
> - to develop ML tools, designed for and tested on planetary science cases submitted by the community, and to provide sustainable, open access to the resulting products, together with support documentation
> - to foster wider use of ML technologies in data driven space research, demonstrate ML capabilities and generate a wider discussion on further possible applications of ML
> - to identify scientific and commercial applications for the ML tools developed through the JRA tasks
>
> **Description of work**
> This JRA will be led by IWF-OEAW, co-led by KNOW, and organised around 6 tasks. It will develop ML powered data analysis and exploitation tools that target a set of representative scientific cases selected from about a dozen proposals for specific applications of ML in planetary science submitted by the scientific user community in the course of proposal preparation. Software developed in the course of the JRA will be open source (Apache License 2.0), thoroughly documented and available via a git service, so that all results can be used freely, and further developed and extended by the community.

1.2 Work Package Beneficiaries and Partners

Apart from the WP lead IWF-OEAW, there are eight beneficiaries contributing to our WP. Table 1 lists the acronyms of the WP beneficiaries as used in the Europlanet 2024 Research Infrastructure (EPN2024-RI) proposal and their corresponding institutions.

Table 1. Work package beneficiaries.

Work Package Beneficiaries	
ACRI-ST	ACRI-ST, France
AOP	Armagh Observatory and Planetarium, Ireland
DLR	Deutsches Zentrum für Luft- und Raumfahrt, Germany
KNOW	Know-Center GmbH, Austria
IAP-CAS	Institute of Atmospheric Physics, Academy of Sciences of Czech Republic, Czech Republic
INAF	National Institute for Astrophysics, Italy
IWF-OEAW	Space Research Institute, Austrian Academy of Sciences, Austria
LMSU	M.V. Lomonosov Moscow State University, Russia
UNIPASSAU	University of Passau, Germany

1.3 Science Cases

The science cases proposed by the planetary science community in the course of proposal preparation are listed in Table 2. The proposal by GMAP covers different cases dealing with the detection and classification of various planetary surface features, as for example mounds and pits.

Table 2. List of science cases.

Proposer	Science Case
IAP-CAS	Detection of plasma boundary crossings at planetary magnetospheres and solar wind
	Classification of plasma wave emissions in electromagnetic spectra
INAF	Mineral identification via reflectance spectra [possible applications foreseen in GMAP]
DLR	Classification of surface composition on the surface of Mercury [resulting data products can be used for GMAP]
AOP	Abundance of asteroids in Earth-like orbits from STEREO images
GMAP	Automatic recognition and analysis of planetary surface features
IWF-OEAW	Detection and classification of CMEs and CIRs in in-situ solar wind data
LMSU	Search for magnetopause/shockwave crossings on Mercury based on MESSEN-GER data

1.4 Deliverables and Milestones

There are eight WP-specific deliverables and three milestones, listed in Table 3. The first milestone, i.e., defining and documenting the requirements for the ML tools, was met on time. The requirements document is designed to be a ’living’ document, which can be updated and modified in the course of the project

Table 3. List of deliverables (D) and milestones (MS).

Abbrevia-tions	Description	Month due	Finished
D1	Annual Report 1	M12	
D2	Annual Report 2	M24	
D3	Tutorial on Machine Learning and Basic How Tos (initial release)	M25	
D4	Demonstrator and Documentation of Data-Processing Techniques	M42	
D5	Demonstrator and Documentation of Time-based Signal Analysis and Automatic Classification Tool	M42	
D6	Demonstrator and Documentation of General Classification Toolset	M42	
D7	Annual Report 3	M36	
D8	Tutorial on Machine Learning and Basic How Tos (final release)	M42	
MS11	Requirements for ML tools documented	M4	?
MS51	ML Demonstrators implemented and tested	M24	
MS86	ML Demonstrators fully validated and integrated	M42	

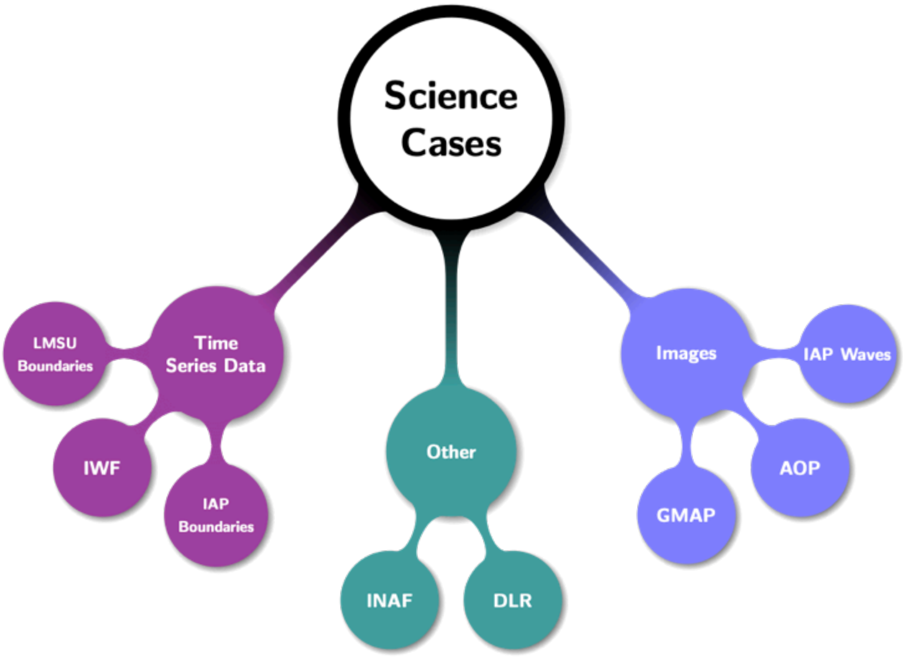


Figure 1. Clusters of science cases according to the corresponding data used.

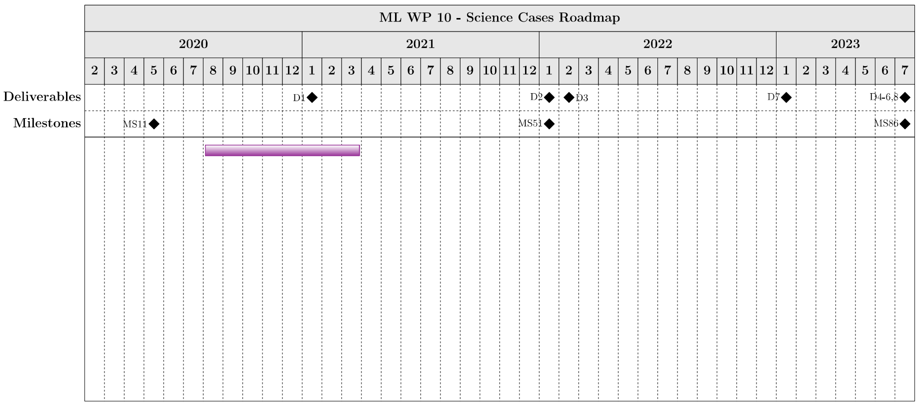
1.5 Explanation of the Work

1.5.1 Preparational Work

To better understand the proposed science cases and the requirements for them, a short questionnaire was prepared by IWF-OEAW and was sent to the WP beneficiaries. The responses to those questionnaires were used to cluster the science cases according to the main type of data used for the cases (see Figure 1). The two cases listed in the cluster ‘Other’ use spectral data, which basically do not fall in one of the other two categories, but both may influence each other and both may benefit from the codes developed for the other cases. The idea behind the clustering is that the science cases within one cluster can be tackled with similar approaches. Thus, the codes/tools developed for one of the cases can be used with (small) modifications for the other cases in the same cluster.
After clustering the science cases, two cases were selected as the first to be developed: the LMSU Boundaries science case from the cluster ’Time Series Data’ and the GMAP Mounds science case from the cluster ’Images’. These two cases form the starting point from which we will then go on to the other cases. Figure 2 shows a roadmap for our science cases, in order to have a guideline for our WP activity. At the end of the first year of the project we are working on four science cases: IWF ICMEs, LMSU Boundaries, GMAP Mounds, GMAP Pits. The first three of these will be brought to an end in February and March 2021 and we will start with the cases IAP Boundaries and AOP Asteroids (see Figure 2).

1.5.2 Infrastructure

UNIPASSAU has set up a private Gitlab group to serve as a development environment for the WP (https://gitlab.padim.fim.uni-passau.de/RP-20-WP10-EPN2024). For each science case, a repository will contain all relevant documents, i.e., descriptions of the science cases, links to publications or other relevant information, data sets (or links to data sets), code scripts, and presentations. Those files are primarily thought to be only for internal WP use. The code scripts on the private repositories are in general not final scripts, but rather in a development state, and are not expected to be used by others than the WP beneficiaries.



deadlines of the WP deliverables and milestones.

However, in addition to that private Gitlab development environment, IWF-OEAW has set up a public GitHub account (https://github.com/epn-ml). In the public GitHub repositories, we will place all the documents and files that are suitable to be made public, e.g., descriptions of the science cases, links to publications and other relevant information, final data sets, working code scripts, and presentations.

A website, our so-called ML Portal (ml-portal.oeaw.ac.at), was installed and serves as an access point to our activity. In the portal, among other content, we will provide:

- An introduction to the ML activity within EPN2024-RI
- General information about ML
- ML tutorials
- Python Jupyter notebooks with different ML tools
- Downloadable ML tools and/or links to them, i.e., Python scripts
- Tutorials on how to use the tools and how to modify them for specific needs
- A list of presentations and publications with results of the ML activity
- Announcements of upcoming events (workshops, sessions at conferences, etc.)

Descriptions of the science cases, general information about ML in EPN2024-RI, links to presentations and announcements of upcoming events are already available on the portal. Right now, we are embedding Jupyter notebooks for ML tutorials and we are showing parts of our first ML tools developed for the first three science cases (LMSU Boundaries, GMAP Mounds, IWF ICMEs).

Just recently, we created a Twitter account, ML Europlanet (@ml_epn), to increase our range to communicate our activities and latest developments to interested parties.

1.5.3 EOSC study

A first draft on how to onboard the services that will be developed as part of the JRA4 into the EOSC has been provided (https://europeansf.sharepoint.com/:b:/r/sites/EuroplanetSociety/Europlanet%202024%20RI/VA/WP10%20Machine%20Learning/EOSC_study_v0.33.pdf?csf=1&web=1&e=NC8SIy). It describes the EOSC, EOSC portal and hub, and the European Grid Infrastructure (EGI). The EOSC is like a catalogue of services that can be accessed via the EOSC portal. By onboarding a service, the service is essentially listed in the EOSC site (like a shop window), but it is hosted by the service provider. The EOSC expects mature services (TRL8/9) to be onboarded, and we will further explore possibilities to onboard our ML demonstrator services on the EOSC. A preliminary list of requirements for onboarding has been identified.

1.5.4 LMSU Boundaries Science Case

The goal of this case is to improve our understanding of Mercury's magnetosphere and its dynamics. As the configuration of the magnetic field and trajectories of charged particles in the magnetosphere are heavily influenced by its geometry, to this end an ML model for automatic detection of magnetopause and bow shock crossings is trained based on MESSENGER magnetometer data. The resulting data set of crossing times and positions is to be used in conjunction with the paraboloid magnetosphere model to compute the magnetic field lines in the magnetosphere; these will subsequently be used to perform modelling of trajectories of particles sputtered from the surface of the planet by space radiation.

The first task was to determine automatically the bow shock and magnetopause crossings. We have manually labelled the data set based on MESSENGER magnetometer measurements averaged over 1 second intervals. To identify bow shocks, we first subtract planetary dipole magnetic field components from the magnetometer measurements, compute the magnitude of the remainder attributed to external sources, apply Savitzky-Golay filter to smooth the time profile of the remainder and compute its second derivative. The first and the last second derivative spikes as determined by z-score are assumed to be the enter and exit bow shock crossings respectively. Magnetopause boundaries were eyeballed using the cartesian components of the magnetic fields in the Mercury Solar Orbital coordinate system. During magnetopause crossings at least one of the components in the magnetogram experiences a sharp growth; the exact component depends on the spacecraft position. The beginning and ending points of this growth region are assumed to determine the magnetopause crossing edges.

We select the spacecraft position coordinates and magnetic field coordinates as input features, and feed these vectors into a Convolutional Neural Network (CNN) based classifier, trained on five classes namely interplanetary magnetic field (IMF), bow shock (BS) crossing, magnetopause (MP) crossing, magnetosheath and magnetosphere. The network comprises of two blocks trained in an end-to-end fashion. The first CNN block extracts shape features from the signals, activations from which are fed into the classification block, comprising of a two layer fully connected network with sigmoid outputs.

The first set of prepared data contained labels from first 50 epochs of the MESSENGER spacecraft. This was split into an 80/20 set of train and validation sets. The validation results of the network trained for first 30 training epochs is expressed below in Figure 3. We report the evaluation metrics has an overall accuracy of 98.99%, with average recall of 96% and a precision of 87% on the BS and MP crossings. In the next steps, we plan to explore optimal thresholding methods to find precise localized timestamps for each boundary crossing.

Preliminary results of this science case were presented at the Europlanet Science Congress (EPSC) 2020. The presentation can be found on our ML Portal and on our Github account.

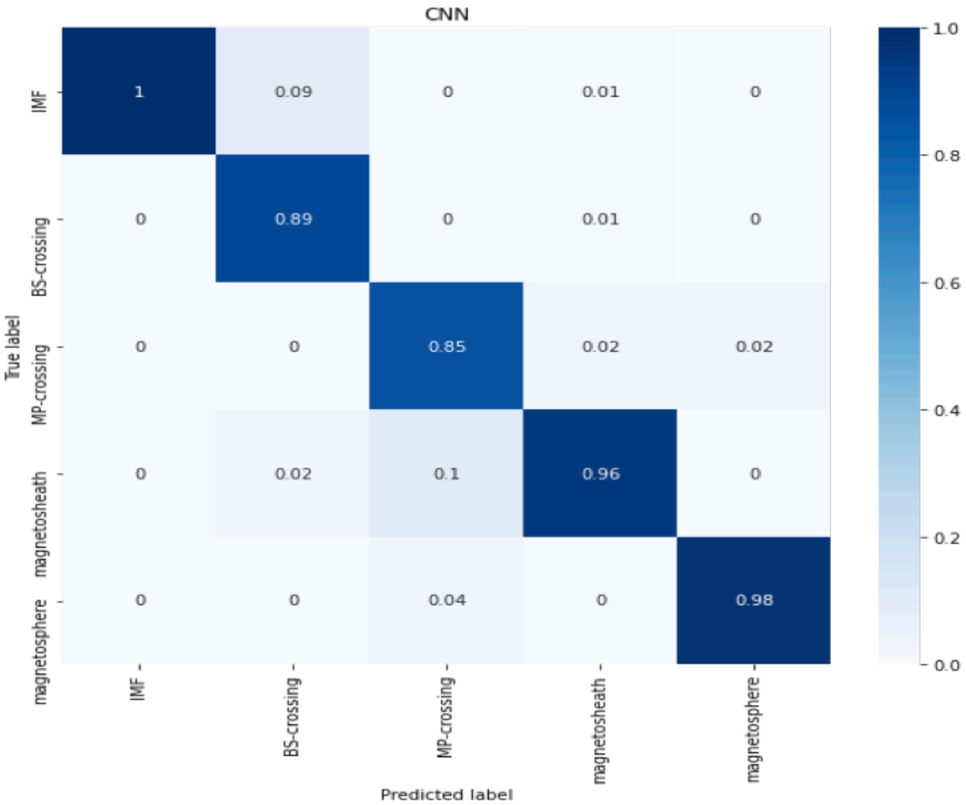


Figure 3: Confusion matrix for the classification performance on boundary crossings. Precision: 87%, Recall: 96%.

1.5.5 GMAP Mounds Science Case

The GMAP Mounds identification science case aims to develop a generalised machine learning pipeline for the localisation and characterisation of specific geomorphological features (mounds) that are present on the surface of Mars. Mounds are positive relief features that can be ascribed to a variety of phenomena (e.g., De Toffoli et al., 2019). They can be related to monogenic edifices due to spring or mud volcanism, rootless cones on top of lava flows, pingos and so on. The focus of the investigation is related to the sedimentary/spring case of mud extrusion or sulphate oversaturated fluids. These objects usually are widespread regionally and/or contained in large complex craters (i.e., tens of km in diameter) often in populations of several hundreds/thousands. Previously, automatic detections were performed in some of these cases (Pozzobon et al., 2019) using topographic data in limited areas (i.e., Digital Terrain Models (DTMs) as rasters whose cells represent height values) in order to discriminate these objects in terms of pre-trained morphometric parameters and map them. Due to the scarcity of high-resolution DTMs and poor area coverage, the ML WP challenge is to reach the ability to detect such mound features by using simple grayscale panchromatic images at mid-high resolution with no need of topographic information.

Keeping the challenge of low training samples in mind, we explored adversarial based deep neural network generators to segment the DTMs into mound and non-mound regions. We condition the GAN with corresponding masks to achieve tractability over generated output modes, and train it using a conditional GAN objective combined with a reconstruction loss. The Generator comprises of a traditional 5-layer U-Net architecture, and the Discriminator is a replica of the encoder part of the Generator without the skip connections. The network is trained for 90 epochs.

Based on preliminary experimentations, the problem is not trivial since the morphological features are not distinctly present in the DTMs, and the training tiles belong to a single DTM. Some exemplary outputs are displayed in Figure 4. Overall, the simple isolated mounds are particularly hard to identify due to overlapping features with non-mound hilly regions.

The network is evaluated quantitatively on the validation set (20%) of the training image, and qualitatively on the test image. The confusion matrix depicting the evaluation performance is illustrated in Figure 5.

In the next steps, we plan to improve the performance by increasing training size with data augmentation. Using the GAN, we are able to generate more samples which we will exploit use to make our classifier more robust and decrease the false positives and negatives. In addition, we will learn the simulation parameters which will allow us to induce domain specific knowledge automatically into the network. This simulation network will serve as foundation for a general terrain simulator that could be adapted dynamically to changing terrains.

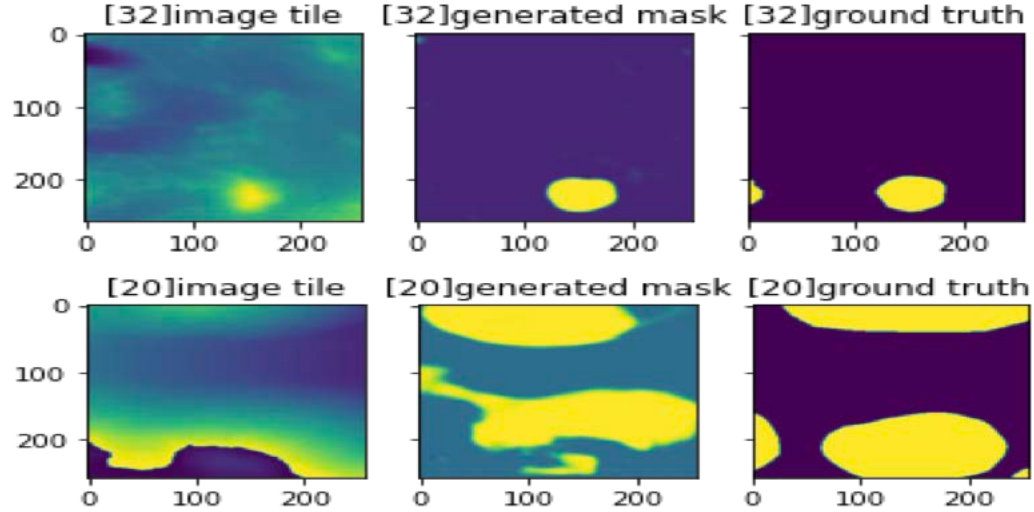


Figure 4: Examples of some network outputs compared with the corresponding ground truth

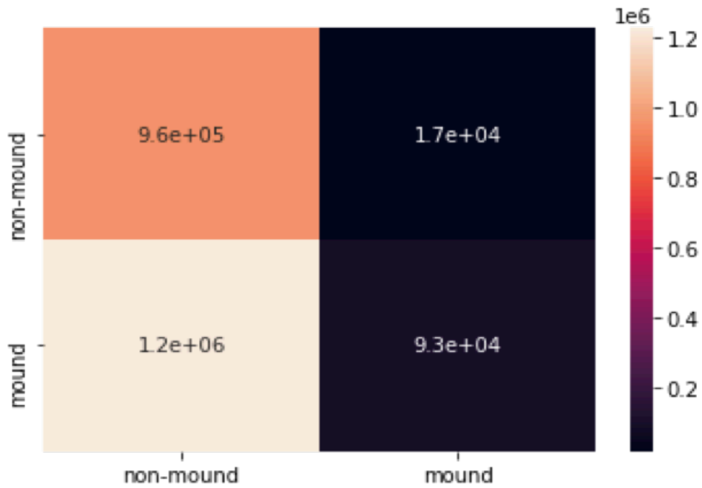


Figure 5: Confusion matrix for mound segmentation performance using the conditional GAN architecture. Mound classification precision: 84%.

In parallel, we will also investigate another line of approach and proceed in an unsupervised fashion, which could, if successful, overcome the problem of limited label availability. The biggest obstacle in this approach is the robustness, which could be improved by introducing suitable features.

References:
Pozzobon, R., et al. (2019), Fluids mobilization in Arabia Terra, Mars: Depth of pressurized reservoir from mounds self-similar clustering, Icarus 321, 938, doi:10.1016/j.icarus.2018.12.023
De Toffoli, B., et al. (2019), Surface Expressions of Subsurface Sediment Mobilization Rooted into a Gas Hydrate-Rich Cryosphere on Mars, Scientific Reports 9, 8603, doi:10.1038/s41598-019-45057-7

1.5.6 IWF ICMEs Science Case

Interplanetary coronal mass ejections (ICMEs) are one of the main drivers for space weather disturbances. In the past, different machine learning approaches have been used to automatically detect events in existing time series resulting from solar wind in situ data (e.g., Nguyen et al., 2019; dos Santos et al., 2020). However, classification, early detection and ultimately forecasting still remain challenging when faced with the large amount of data from different instruments. While CNNs are often used to discover objects or patterns in images or data series, there are two main problems when facing our specific task: high duration variability and a rather ambiguous definition of start and end time. The first step in this science case was the reimplementation of a model proposed by Nguyen et al. 2019, which had previously been tested on WIND data and achieved a maximum recall and precision of around 84%. It was additionally tested on STEREO-A and STEREO-B data, which contain less variables than WIND data. The model was still able to achieve similar recall on all three data sets, but the precision went down to around 30%. Furthermore, the preciseness when delivering start and end times is limited. The next steps in this science case are the alignment of different datasets in order to process more training data for a combined model and the inclusion of a more advanced post processing step to improve performance.

Preliminary results of this science case were presented at the European Space Weather Symposium (ESWS) 2020. The presentation can be found on our ML Portal and on our Github account.

References:

Nguyen, G., et al. (2019), Automatic Detection of Interplanetary Coronal Mass Ejections from In Situ Data: A Deep Learning Approach, Astrophys. J. 874, 145, doi:10.3847/1538-4357/ab0d24
dos Santos, L.F.G., et al. (2020), Identifying Flux Rope Signatures Using a Deep Neural Network, Sol. Phys. 295, 10, doi:10.1007/s11207-020-01697-x.

1.5.7 GMAP Pits Science Case

A deep learning object detection algorithm YOLO has been applied to detect pit craters on the surface of Mars from images obtained by the HiRISE instrument on board the Martian Reconnaissance Orbiter (MRO). Current performance reaches 74% recall and 70% precision. Image augmentation and expanded labelling were shown to provide better performance results. The extraction of statistically relevant products was also demonstrated.
Figure 6 shows an example of pit detection (inference) with a confidence threshold of 0.5 on a test image using the pre-trained, augmented configuration trained on an expanded dataset. This resulted in high confidence predictions on 15 true positives (blue boxes) and 0 false positives, but the network fails to detect 3 small pits in the image (red boxes).

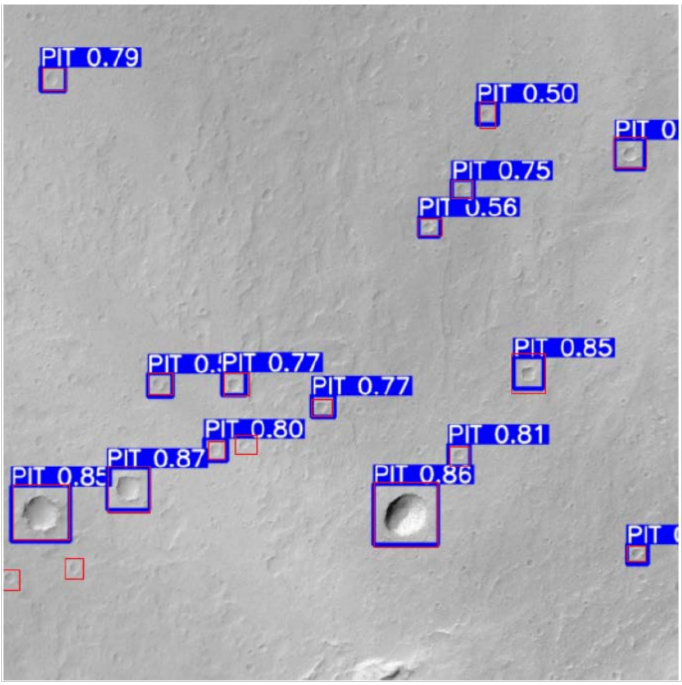


Figure 6. Example of pit detection (inference) on a test image.

1.5.8 Other Science Cases

The data sets for the other science cases will be prepared in time. For example, for the AOP asteroids science case, which is one of the next science cases being worked on, information material, presentations and data sets are available on GitLab and public material is available on GitHub.

For the DLR case, we are already integrating the input data set of Mercury surface spectroscopic data in VESPA.

2. Updated Dissemination Plan

Due to the COVID-19 situation, we are now preparing to give online workshops instead of face-to-face workshops. The first ML tools, i.e., Python scripts developed for the first three science cases (LMSU Boundaries, IWF ICMEs, GMAP Mounds), will be available for the community in spring 2021, when we will also conduct our first workshop. Part of the scripts will be embedded in our ML Portal to provide a quick possibility to get acquainted with the code. The full scripts will be available through our public GitHub account.

Prior to the workshops, we will provide tutorials for registered participants on how to install software, packages, etc. necessary to run the Python scripts and on how to use the tools. At the time of the workshops, the participants should have a functional environment and should be familiar with the tools. They also should know what they want to do with the ML tools and how they want to use them for their specific science cases. During the workshops, we will answer questions about the ML tools and we will provide support to modify them.

3. Updated Data Management Plan

We will update our Data Management Plan (DMP) to clarify issues raised in a comment by the VA Board Review (see Section 4.1).

4. Comments and Suggested Actions from VA Board Review

The external VA Review Board provided a review report about the VA activities. We want to underline that WP 10 is not a VA, but a JRA, thus it cannot be reviewed in the same manner as the other VAs. In the following, we list the comments and suggested actions and our replies.

4.1. Comments

> The goal of this project, as expressed in the work package, is to develop ML tools and provide access to the resulting products and documentation. It is not clear, however, what constitutes a 'tool' vs. a 'product' in the ML case in either the work package description or on the current website. For example, is it the intention of the project to distribute trained models for off-the-shelf use, model code for users to adapt to their own use cases, both, or neither? What does it mean to 'integrate ML tools' into the VESPA, GMAP, and SPIDER websites? Defining and describing the goals of the project in plain, non-specialist terms will greatly enhance efforts to raise ML visibility and to encourage new users to consider ML tools for possible solutions.
>
> Apart from setting up the portal website with its introductory text, there is no indication that any work has begun at all on the various deployments and integrations included in the VA task for this project. At this early stage working deployments are not expected, but a schedule and milestones would be, along with an inventory of any significant development that might be needed for deployment in each case. For example, if the intention is to make trained models available through VESPA, would these models be considered a new data product that would require integration into the Virtual Observatory data model?
>
> Issues of support and standardisation are not addressed in any of the available documentation. Without these, dissemination of new tools and models would be difficult, if not impossible. If there is no community standard, will this project define its own? Whether the intent is to distribute either models or code, there should be a corresponding Data or Code Management Plan to define requirements and expectations.

A DMP for the ML activity does already exist and was included in the DMP for the VAs. It seems that the review board did not have access to the DMP. We defined the term 'data' as used for our WP in the DMP. We also described what will be linked to VESPA and how this will be done. However, we will update the corresponding parts in the DMP to better clarify the raised issues.

> The GitLab repository has been established, but it requires registration to access. And while it has been populated, it does not reflect the sort of commit and update activity one would generally expect for an active development site. Consequently, it is not clear whether the GitLab is being employed as a backup, as a distribution site for registered users, or as a development site.

As explained in Section 1.5, the GitLab group is a private group intended for internal WP development, and documentation. We set up a public GitHub account to provide an open environment, where we put public documents and files to be used by the community.

> The version number of the European Open Science Cloud (EOSC) integration study would indicate that the document is not complete. There are some significant topics missing, including whether GPUs and sufficient storage resources are available from the EOSC, or what alternatives might be available if the EOSC integration should not prove feasible.

The EOSC study was intended to be a preliminary study to have a quick look at EOSC, what it is and how we could in principal onboard our tools. The study is not complete, as mentioned in the comment. However, the study already shows that EOSC does not provide computing resources, but it is rather a 'market place', where we can advertise our tools and where the user is re-directed to resources of the WP beneficiaries.

> While disseminating information about the concepts and uses of machine learning to the planetary science community is a core part of the work plan, there seems to have been no progress on this to date. The EPSC and ESWS presentations reviewed did not provide any background or educational material describing ML methods.

The presentations at EPSC and ESWS were scientific presentations, focussing on the scientific results of the science cases, rather than aiming at educational ML purposes. The first tools from the first three science cases will be finalised and available for the community in spring 2021. In the first half of 2021 we are planning to hold a workshop to provide support on how to use and modify these tools. In the second half of 2021 we are planning to hold another workshop with the same aim, probably having already more tools available. Furthermore, now that we have preliminary results of our first science cases, we will provide tutorials on our ML Portal about the approaches and algorithms used.

4.2. Suggested Actions

> Prioritise development of the website as the primary introduction to ML for new or existing planetary science users. Introductory documentation and plain-language descriptions of planned development and potential applications in planetary science are essential to creating and maintaining interest in the project from both existing and new users.

> Set and publicise (on the website, for example) schedules for development and deployment of the planned tools. Including high-level descriptions of how these tools could be applied, with expected outputs or benefits, would further encourage interest in ML development and deployment.

The ML Portal went online at the end of October 2020. We are continuously adding more and more content. We will take up the suggestion of providing introductory documentation and descriptions of our ML tools and their potentials in planetary sciences on the portal as soon as possible.

> Develop specific work plans (along with schedules) for integration of ML tools into the GMAP, SPIDER, and VESPA portals.

We will develop work plans and schedules for these tasks.