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 D10.2

Deliverable Title:	ML 2nd Year Report
Due date of deliverable:	31st January 2022
Nature[1]:	Report
Dissemination level[2]:	Public
Work package:	WP10
Lead beneficiary:	IWF-OEAW
Contributing beneficiaries:	KNOW, UNIPASSAU, ACRI-ST, DLR, INAF, IAP-CAS, AOP
Document status:	Draft

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

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)

Executive Summary / Abstract:

This annual report summarizes the work done in Work Package 10 'Machine Learning Solutions for Data Analysis and Exploitation in Planetary Sciences' during the second 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 code developed for specific science cases. Work package 10 is organized around six tasks that target management and coordination of the activities, the development of machine learning based data analysis code and the dissemination of the tools as well as integration of the results into VESPA, GMAP and SPIDER where appropriate. Despite delays in the development work due to the ongoing Covid-19 pandemic, work on all of the six tasks has been progressing. Developments on three science cases are considered to be finished and work on four other science cases is progressing. We conducted two workshops at the EPSC 2021 conference, introducing two of our machine learning pipelines. We put up first tutorials on our Machine Learning Portal as well as on our public GitHub repositories. ML organized machine learning sessions at EGU21 and EPSC2021, and had presentations at many conferences (LPSC2021, EGU21, EPSC2021, ESWW 2021, AGU Fall Meeting 2021). We started collaborations with national (FWF project at IWF) as well as international (EU Horizon 2020 EXPLORE project) research projects, and started a series of fireballs workshops together with NA2. An EPN-TAP server was set up at the IWF, on which we started to integrate first data sets of our science cases into VESPA. Furthermore, first steps were undertaken to include our pipelines in SPIDER.

Table of Contents

Table of Abbreviations	
AGU	American Geophysical Union
ASPP	Atrous Spatial Pyramidal Pooling
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
EPN-2024-RI	Europlanet 2024 Research Infrastructure
EPSC	Europlanet Science Congress
ESA	European Space Agency
ESWW	European Space Weather Week
GAN	Generative Adversarial Network
GMAP	Geologic MApping of Planetary bodies
GPU	Graphics Processing Unit
HCS	Heliospheric Current Sheet
ICA	Independent Component Analysis
ICME	Interplanetary Coronal Mass Ejection
IMF	Interplanetary Magnetic Field
IoU	Intersection over Union
JpGU	Japan Geoscience Union
JRA	Joint Research Activity
LPSC	Lunar and Planetary Science Conference
LSTM	Long-Short-Term Memory Network
MASCS	Mercury Atmospheric and Surface Composition Spectrometer
MESSENGER	MErcury Surface, Space ENvironment, GEochemistry, and Ranging
ML	Machine Learning
MP	Magnetopause
MS	Milestone
NA	Networking Activity
PMC	Project Management Committee
SDA	Scientific Data Application
SPIDER	Sun Planet Interactions Digital Environment on Request
TRL	Technology Readiness Level
UMAP	Uniform Manifold Approximation and Projection
VA	Virtual Access
VESPA	Virtual European Solar and Planetary Access
WP	Work Package

1 Explanation of WP10 Work & Overview of Progress

a. 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.

Work Package Beneficiaries

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

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 MESSENGER data

Deliverables and Milestones

There are nine deliverables and three milestones for our WP, listed in Table 3. All milestones and deliverables were met in due time.

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

Abbreviations	Description	Month due	Finished
D10.1	Annual Report 1	M12	✓
D10.2	Annual Report 2	M24	
D10.3	Tutorial on Machine Learning and Basic How Tos (initial release)	M31	
D10.4	Demonstrator and Documentation of Data-Processing Techniques	M42	
D10.5	Demonstrator and Documentation of Time-based Signal Analysis and Automatic Classification Tool	M42	
D10.6	Demonstrator and Documentation of General Classification Toolset	M42	
D10.7	Annual Report 3	M36	
D10.8	Tutorial on Machine Learning and Basic How Tos (final release)	M42	
D10.9	Annual Report 4	M48	
MS11	Requirements for ML tools documented	M4	✓
MS51	ML Demonstrators implemented and tested	M30	
MS86	ML Demonstrators fully validated and integrated	M42	

b. Explanation of the work carried in WP

Task 1 - Management and Coordination

This task oversees the management of the ML JRA4, coordinates the activities within the WP and with the other WPs and reports to the PMC.
We updated our science cases roadmap (see Figure 1) from last year to account for some delays due to the Covid-19 pandemic and science case specific problems, e.g., bad quality of data, trying out different approaches, etc.

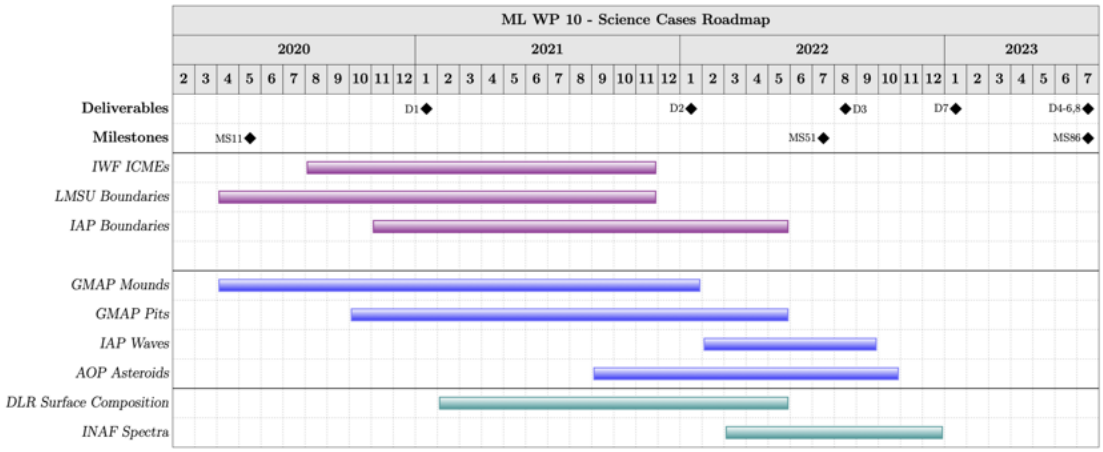

Figure 1: Timeline for the science cases. Also shown are the deadlines for the WP deliverables and milestones.

Task 2 - Requirements for Machine Learning, Tool Validation and Communication

Infrastructure

The three platforms set up during the first year of the WP, namely the (private) GitLab group, the GitHub organization and the ML Portal, were further maintained. There is now more information about our activity on the ML Portal, e.g., more information about the science cases, presentations, news regarding ML conferences and sessions and tutorials.

Presentations and Workshops

A presentation introducing the ML WP and its activities was given at the Open Planetary Lunch in June 2021 as well as at the AGU21.

Fireball-tracking networks around the world are assisting in the recovery of fragments of fresh meteorites and understanding where in the solar system they originated. In collaboration with NA2, the ML WP organised a workshop on 11-12 June 2021 to bring together observers from different fireball networks, along with ML experts, to discuss how ML can support the fireballs community and to advise on handling the data collected. This workshop was the first in a series of four - the next will take place (again virtually) on 4-5 February 2022.

Two ML pipelines have been presented in two workshops during EPSC2021 - the pipeline for the IWF ICME science case as well as the one for the LMSU boundaries science case. Both pipelines are available on GitHub.

Presentations with results of the science cases are mentioned in the section about the individual science cases.

Collaborations

We started a collaboration with two research projects at the IWF. Out of one of these collaborations a publication about the prediction of the magnetic field Bz component of ICMEs arose:

Reiss, M., et al., (2021), Machine Learning for Predicting the Bz Magnetic Field Component From Upstream in Situ Observations of Solar Coronal Mass Ejections, https://doi.org/10.1029/2021SW002859.

Further, we started a collaboration with the EU Horizon 2020 project EXPLORE. On the one hand, ML supports their Lunar Data Challenge. On the other hand, we are investigating the possibility to integrate our ML pipelines into the EXPLORE platform (see description of Task 6).

Task 3 - Data Pre-Processing, ETL and Feature Engineering

The aspects of data pre-processing and feature engineering are covered in the descriptions of the work for the individual science cases. Most science cases thereby utilize standard pre-processing methods or work on the raw data through end-to-end learning. However, we also explore new routes to automate pre-processing. For example, the GMAP Mounds science case utilizes data augmentation in the form of generative adversarial networks to overcome data sparsity. Details on the pre-processing conducted can be found below.

Task 4 - Time-based Signal Analysis and Automatic Classification

IWF ICME 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. 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.

After the reimplementation of a model proposed by Nguyen et al. (2019) in year 1 of this WP, the model was tested on STEREO-A and STEREO-B data as well as on WIND data. All three contain less variables than the original data set used by Nguyen et al. At a similar recall as for the original set, the precision for all three datasets was only around 30% and the accuracy in delivering start and end times was limited.

The next step was to align all three data sets in order to process more training data for a combined model. It was tested on held out datasets for WIND, STEREO-A and STEREO-B. Surprisingly, this did not sufficiently improve performance and lead us to explore other approaches.

Starting from the reimplementation a post processing step based on YOLO v5 (ultralytics) was investigated, in order to improve performance. Even though first results seemed promising, the idea was later discarded due to unsatisfactory results and the laborious pipeline. Since the ultimate goal is an explicit and widely applicable pipeline, it was decided to abandon the general approach of using multiple basic neural networks and the similarity measure used by Nguyen et al. (2019) completely and compose it as a segmentation problem instead.

We proposed a pipeline using a UNet (Ronneberger et al., 2015) including residual blocks, squeeze and excitation blocks, Atrous Spatial Pyramidal Pooling (ASPP) and attention blocks, similar to the ResUNet++ (Jha et al., 2019), for the automatic detection of ICMEs. Comparing it to last year's results, we find that our model outperforms the baseline regarding GPU usage, training time and robustness to missing features, thus making it more usable for other data sets, as well as the three aligned data sets. The confusion matrix is shown in Figure 2.

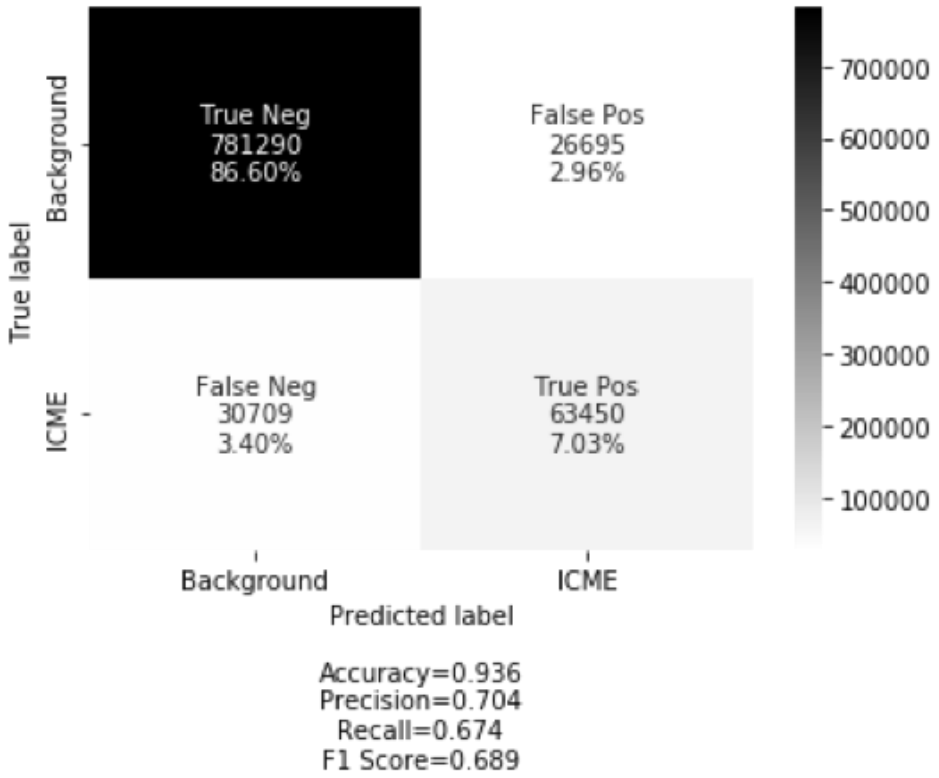

Figure 2: Confusion matrix for the results of the ICME science case.

The relatively fast training allows straightforward tuning of hyperparameters. Our proposed pipeline can be used for any time series segmentation problem. The straightforward implementation allows a simple extension to a multiclass classification problem and paves the way to include corotating interaction regions into the range of detectable phenomena within our pipeline. Furthermore, we hope to apply our model to similar problems in the future.

Results of this science case were presented at the EGU21, at EPSC2021, at ESWW 2021, and at AGU21 (see presentations on the ML Portal and on GitHub). A further presentation was given in May 2021 at an international working group called 'CMEs, CIRs, HCS and large-scale structure' (led by, among others, Christian Möstl and Silvia Perri). This ML pipeline was presented in a workshop at EPSC2021 and is, together with a tutorial, available on our GitHub repository. A publication will be submitted soon.

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
Jha, D., et al. (2019), Resunet++: An advanced architecture for medical image segmentation, arXiv e-prints, arXiv:1911.07067

LMSU Boundaries Science Case

The goal of this case is to improve our understanding of Mercury's magnetosphere and its dynamics. We utilise the data recorded by the MESSENGER (MErcury Surface, Space ENvironment, GEochemistry, and Ranging) spacecraft, which collected vast amounts of heterogeneous data during its approximately 4000 orbit voyage, most interestingly the magnetic field data from the magnetometer. A typical orbit involved passing from the interplanetary magnetic field through the bow shock, the magnetosheath, the magnetopause, the magnetosphere of Mercury, and thereupon the same sequence in reverse. Since a mercurial year is about 88 Earth days, several years' worth of magnetometer data was recorded. This is nice because several variations in environmental configurations are recorded, which is useful to build automatic models for event recognition. 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 can subsequently be used to perform modelling of trajectories of particles sputtered from the surface of the planet by space radiation.

Based on data from the mission, several global models of the magnetosphere were proposed (e.g., Winslow et al., 2013; Philpott et al., 2020). However, they could only describe an average shape of the bow shock and magnetopause crossings and can be prone to missing the statistical nuances in the data. Given large data, neural networks can be expected to approximate complex functions, which often surpass deterministic and rule-based methods, in a variety of time series tasks like classification (Fawaz et al., 2019), time series forecasting (Lim and Bohren, 2021), and rare time series event detection (Nguyen et al., 2018). We leverage these to develop a predictor that can be used in real-time during orbit to predict magnetic region for each step in a short window of observation. Figure 3 illustrates the different crossing labels for an exemplary orbit.

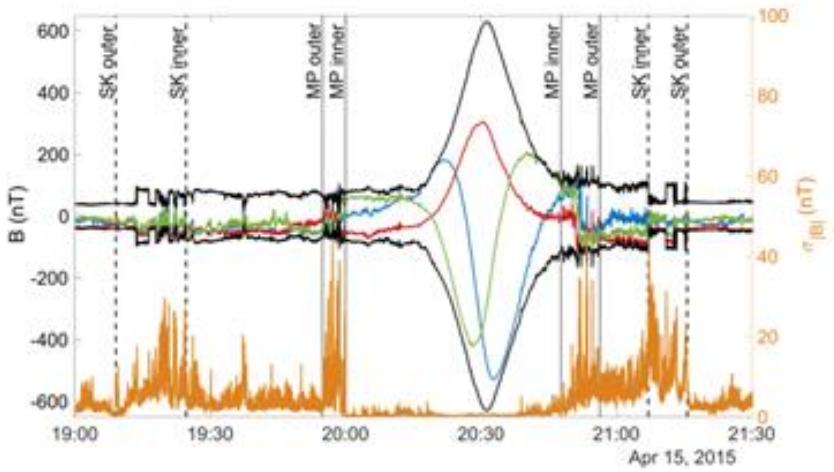

Figure 3: Exemplary labelled orbit from the work of Philpott et al. (2020).

The use of statistical neural networks allows us to explore another aspect: with the help of active learning, it is possible to add samples to the training process incrementally. With this, we can examine how the model scales its predictive capacity with increasing data, and thus study how the variations such as changing solar wind and environmental conditions affects the manifestation of boundary signatures. To begin with, different orbits can be expected to have some element of similarity in the magnetic field structure, yet would have large variations in the same segments at different conditions. It is also interesting to study what the minimum amount is for the data needed to be able to generalise these phenomena for future missions such as BepiColombo.

The dataset was manually labelled with the boundary crossings. To identify bow shocks, we first subtracted planetary dipole magnetic field components from the magnetometer measurements, computed the magnitude of the remainder attributed to external sources, applied the Savitzky-Golay filter to smooth the time profile of the remainder and computed 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. To supplement these, we also used the boundaries marked by Philpott et al. (2020) for a few orbits.

The distribution of the different magnetic regions, after annotation, is reported in Table 4. The boundaries of critical interest - bow shock and magnetopause - are minorities with only 3.7 and 2.3 % representation. The table highlights the data imbalance issue that requires investigating special techniques to ensure the predictor does not bias towards the overrepresented classes.

Table 4. Class-wise distribution present in the data.

Label	Magnetic region	Statistical distribution
0	Interplanetary magnetic field (IMF)	65.4 %
1	Bow shock crossing (SK)	3.7 %
2	Magnetosheath (MSh)	14.5 %
3	Magnetopause crossing (MP)	2.3 %
4	Magnetosphere (MSp)	14.1 %

As a first step in pre-processing, feature selection was performed to assess the contribution of available features in the estimation of the output. Based on statistical correlations, the magnetic flux features (BX_MSO, BY_MSO, BZ_MSO), spacecraft position coordinates (X_MSO, Y_MSO, Z_MSO) and planetary velocity components (VX, VY, VZ) were found to be most informative. In addition, three meta features namely EXTREMA, COSALPHA and RHO_DIPOLE were selected.

In the feature preparation stage, a sliding window of variable sizes (3 seconds to 3 minutes) with a hop size of 1 second was computed on the time series signal to obtain feature vectors. Finally, the features were normalised to have mean of 0 and a standard deviation of 1. No other pre-processing or engineering was applied in order to allow the deep learning model to engineer features implicitly.

The windowed features are fed first into a block of 3 Convolutional layers with 1D filters, each followed by Batch Normalisation and ReLu activations. The activations obtained at the end of the CNN block are then passed to the Recurrent block with two layers of LSTMs. The final activations are then passed to a fully connected layer with softmax activations. The objective function used for training is Categorial cross entropy, with Adam optimizer.

The sample results in Figures 4 and 5 are from a model trained with two Mercury years of data, which is about 300 orbits.

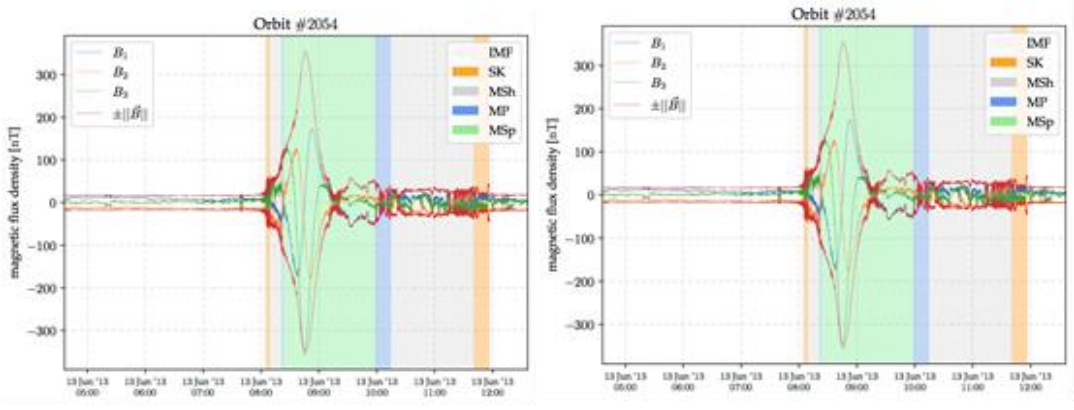

Figure 4: Left: Prediction: right: Ground truth.

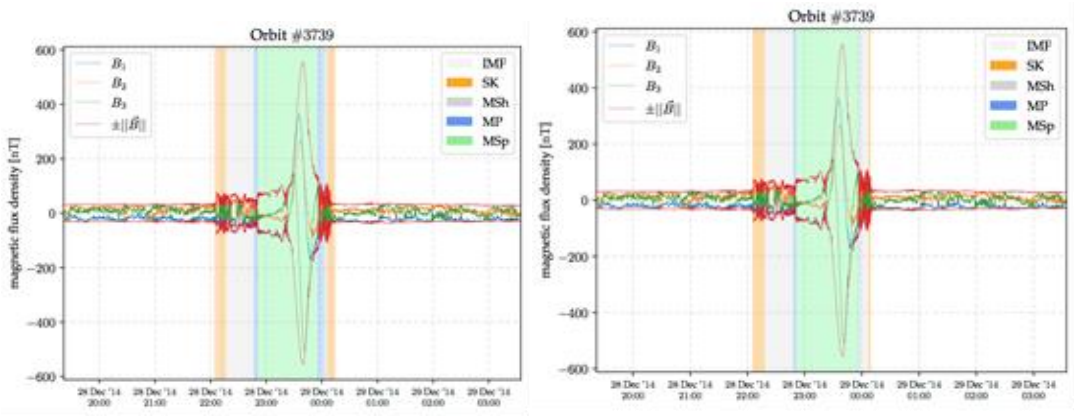

Figure 5: Left: Prediction: right: Ground truth.

The window size used in these experiments is 30 seconds. Overall, the predictor achieves a macro F1 score of about 80% on the bow shock and the magnetopause crossings on a randomly sampled test of 300 orbits. None of the orbits overlap in the train and test sets.

The results from the active learning experiment are still not complete. We are currently in the process of documenting them and we will put them forth in a publication soon.

Results of this science case were presented at the EGU21 as well as at EPSC2021 (see presentations on the ML Portal and on GitHub). This ML pipeline was presented in a workshop at the EPSC2021 and is available on our GitHub repository. A publication will be submitted soon.

References:

Philpott, L.C., et al. (2020), The Shape of Mercury's Magnetopause: The Picture From MESSENGER Magnetometer Observations and Future Prospects for BepiColombo, *J. Geophys. Res. (Space Physics) 125*, doi: 10.1029/2019JA027544

Winslow, R. M., et al. (2013), Mercurys magnetopause and bow shock from MESSENGER Magnetometer observations, *J. Geophys. Res. (Space Physics) 118*, 10.1002/jgra.50237

Fawaz, H.I., et al. (2019), Deep learning for time series classification: a review, *Data Mining and Knowledge Discovery 33,* doi: 10.1007/s10618-019-00619-1

Lim, B., and Zohren, S. (2021), Time-series forecasting with deep learn- ing: a survey, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379*, doi: 10.1098/rsta.2020.0209

Nguyen, V., et al. (2018), Applications of Anomaly Detection Using Deep Learning on Time Series Data. *In: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), doi:10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00078.*

IAP Boundaries Science Case

Planetary magnetospheres create multiple sharp boundaries such as the bow shock, where the solar wind plasma is decelerated and compressed, or the magnetopause, a transition between solar wind field and planetary field. The boundaries are identified by a discontinuity in magnetic field, plasma density, and in the spectrum of high-frequency waves. These measurements are available on many planetary missions, such as Cluster or THEMIS (Figure 6). Due to the high amount of available data, a deep learning approach was found to be well suited to automatically identify the said boundaries. In 2021, we have compiled a large dataset of data collected by several instruments on the ESA Cluster satellites of more than 2000 bow shock crossings encountered between 2001 and 2014 (the crossings were identified visually by humans to prepare this training dataset). The data has been pre-processed and the process of model development has been started. The code to process the original spacecraft data is available on GitHub.

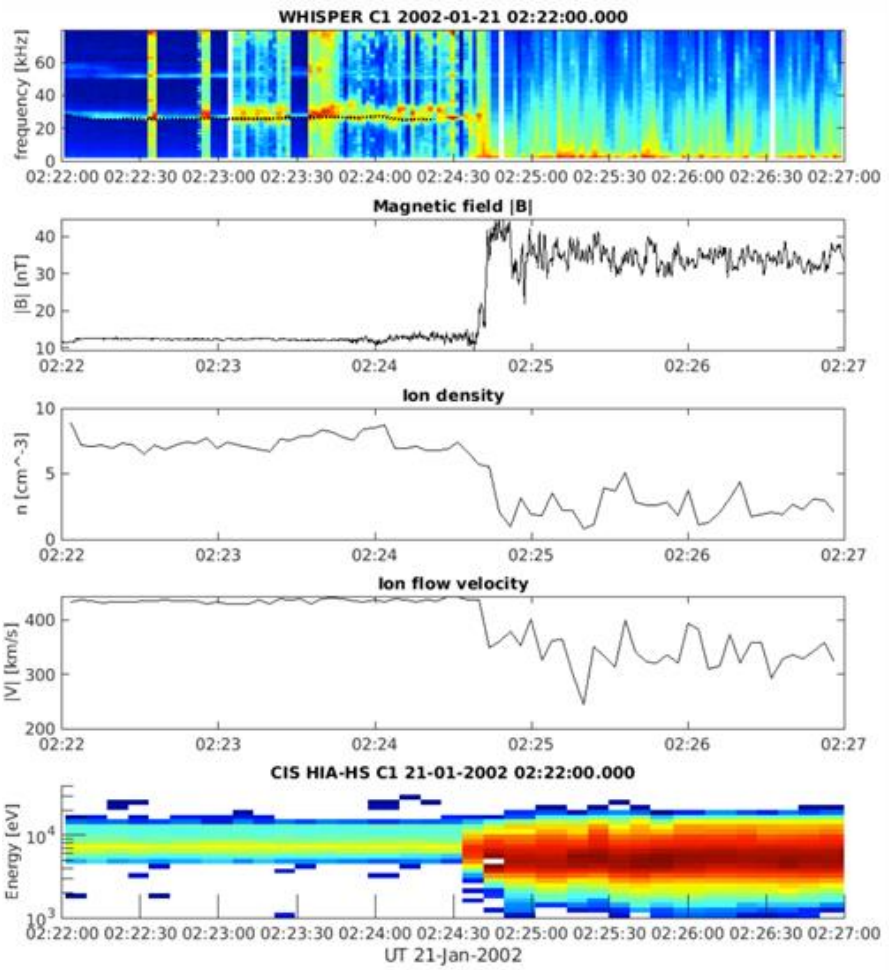

Figure 6: Example of a bow shock crossing.

This science case was presented at the EPSC2021 and the AGU21 (see presentations on the ML Portal).

Task 5 - Images and Other (General) Classification Tools

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 are usually widespread regionally and/or contained in large complex craters (i.e., tens of km in diameter) often in populations of several hundred/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.

The training set consists of two DTMs, one used for training and the other for testing. In the first step, the training DTM is tiled into several smaller fixed sized images. The label masks are created based on the available ground-truth shape files. The images are then scaled to be in range [-1,1]. The training set is then split further into train and validation sets with an 80/20 ratio. The train set is augmented in the next step with image manipulations such as flipping, rotation, rescaling and so on to create a large training set for the segmentation task.

For the initial image segmentation task, a standard UNet (Ronneberger et al., 2015) is trained using the training set. A mean IoU (Intersection over Union) of about 60 % on the validation set is obtained. This result is consistent with another GAN based model, indicating a saturation in information present in the training set.

Due to the limited number of samples to train from, we learn a Generative model (Goodfellow et al., 2020) to approximate the true distribution of the landforms. We generate an augmented set using this approach and train the image segmentation again, observing an improvement of about 10% in the IoU. This is an interesting result, as it indicates that the model can be used to simulate the mound terrains. The approximated distribution space should be then factorisable into a set of independent mechanisms, which could control factors of variation.

A simulator of such likes can be used for controlled generation. Another advantage of latent space learning is that it can offer benefits in downstream tasks, which is an added advantage for storage and efficient searching. We have developed this simulator and we plan to disseminate the method as a publication in the coming months.

Results of this science case were presented at the EGU21 (see presentations on the ML Portal). The ML pipeline is available on our GitHub repository.

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
Ronneberger, O., et al. (2015), U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (eds) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28
Goodfellow, I., et al. (2020), Generative adversarial networks, *Communications of the ACM 63,* doi:10.1145/3422622

GMAP Pits Science Case

To improve the results obtained by the first tool for automated mapping of pits (DeepLandforms-YOLOv5, https://github.com/epn-ml/DeepLandforms-YOLOv5), a change of architecture was necessary. The results obtained by that tool, despite their high quality, need further processing since they are not immediately usable for proper mapping as they are composed only by a pair of coordinates that localize the centre of the detected features. Such detections still need to be properly mapped as polygonal shapes by users. Since this is a highly time-consuming and tedious task, it led to the development of a new tool based on Deep Learning Instance Segmentation, to retrieve not only point coordinates of the detected features, but also a polygonal shape. The obtained results were then compared to the results obtained with the previous tool and with the MGC^3 database (Cushing et al. 2012, 2015), showing good results. A publication and this new tool will be released soon.

Results of this science case were presented at the LPSC2021 (see presentations on the ML Portal).

References:
Cushing, G., et al. (2012), Candidate Cave Entrances on Mars, *J. Cave Karst Stud. 74*, 33–47, doi:10.4311/2010EX0167R
Cushing, G.E., et al. (2015), Atypical Pit Craters on Mars: New Insights from THEMIS, CTX, and HiRISE Observations, *J. Geophys. Res. (Planets) 120*, 1023–1043, doi:10.1002/2014JE004735
Nodjoumi, G., DeepLandforms-YOLOv5. Available online: https://zenodo.org/record/4430015 (accessed on 15 December 2021).

DLR Surface Composition Science Case

In this science case, Mercury surface reflectance data from the MASCS instrument onboard the NASA/MESSENGER mission is analysed. First, NASA/PDS data is converted in a relational DB (PostgreSQL). Then the data is regridded with custom Postgis/PostgreSQL spatial queries. This produces a global hyperspectral data cube image of normalized MASCS visible (VIS) detector spectra, from the first Earth year of

the orbital mission. The cube contains some anomalies, in regions of low coverage or from high levels of spectral variation within a single pixel. Thus, data artifacts, instrumental and photometric residual effects are all removed. The resulting data cube has several hundred features that are compressed via blind signal demixing with Independent Component Analysis (ICA). Initial results show that four components reconstruct the original dataset within the measurement estimated error. The four features were embedded in a two-dimensional space via Uniform Manifold Approximation and Projection (UMAP). No significant small-scale morphology was found after exploring UMAP hyperparameters. Finally, the 2D maps were partitioned with hierarchical agglomerative clustering. Dendrogram gap analysis shows a big gap between data partition in three and four clusters, and three clusters have been chosen as a significant data segregation. At this initial stage, the existence of two large and spectrally distinct regions have been found, which have been designated the polar spectral unit and the equatorial spectral unit (see Figure 7).

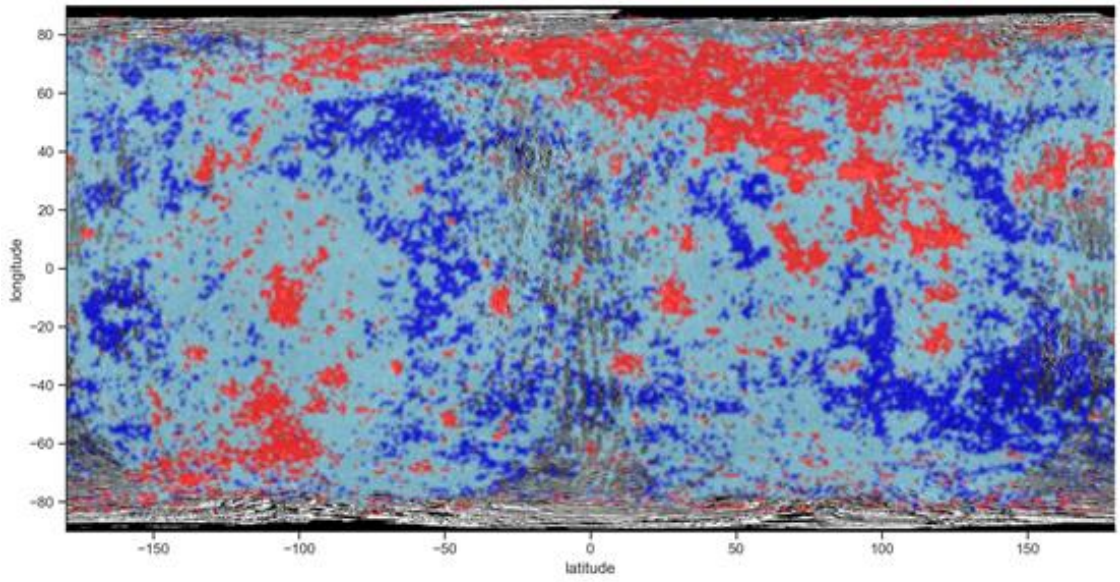

Figure 7: Agglomerative Clustering 3 classes.

The spatial extent of the polar unit in the northern hemisphere generally correlates well with that of the northern volcanic plains and partially to the surface highest temperature models in the equatorial region. This may indicate an interaction between mineral composition and structure and surface temperature, because Mercury reaches a diurnal temperature above 700 K. Chemical data spatial distribution from X-ray and Gamma ray spectrometers show no apparent correlation with the clusters. This could indicate that chemical composition produces no distinctive mineral phases for the instrument or that those phases were altered enough to be indistinguishable by the harsh space environment around Mercury. Further analysis indicates the presence of smaller sub-units that lie near the boundaries of these large regions and may be transitional areas of intermediate spectral characters.

First results of the science case were presented at EGU21 (see presentations on our ML Portal).

AOP Asteroids Science Case

The goal of this science case is to search for asteroids thought to exist along the Earth's orbit that may be leftover material from the formation of our planet. These asteroids always appear close to - or even behind - the Sun in the sky and are therefore difficult to detect from Earth. Images taken from the two STEREO probes which have been studying the Sun and its vicinity since 2009 will be used as the basis for the science case. The spacecraft have been slowly drifting along the Earth's orbit and are able to image the sky from different vantage points around the Sun. This enables the abundance of asteroids in Earth-like orbits to be constrained, including any large (hundred-metre to kilometre size) objects in unstable paths that are not picked up by surveys and present a long-term impact hazard to our planet.

Development in this science case is still ongoing. After first inspection of the corresponding data, we came to the conclusion that we have to reconsider our approach. We might also need to refine and/or re-define the science case.

Task 6 - Virtual Access and Interfaces

The Machine Learning Portal provides the public point of entry to our ML activities. A first draft of how JRA4 services can be onboarded into the EOSC has been provided in year 1 of the WP, including a description of the EOSC, EOSC portal and hub, and the European Grid Infrastructure (EGI). Onboarding a service into EOSC means that the service is listed in the portal of the EOSC site (like a shop window) but is hosted by the service provider. The EOSC expects mature services (TRL8/9) to be onboarded. Further possibilities to onboard ML demonstrator services on the EOSC are being explored. A preliminary list of requirements for onboarding has been identified.

Given the high level of TRL needed to onboard a tool/service to the EOSC, this approach might not be feasible. Thus, we are looking for alternatives.

The EXPLORE platform (https://explore-platform.eu) is a development platform whose main purpose is to validate, test and demonstrate the scientific data applications (SDAs) being delivered by the EXPLORE project. These SDAs will subsequently be deployed also on other platforms – when these are ready – such as ESA Datalabs and ESCAPE SAP. This portability is key to bring the SDAs close to the data.

A joint effort between Europlanet RI 2024 and EXPLORE is now ongoing to update the EXPLORE platform to allow the deployment of JupyterLab-type applications (a technical update is necessary to run JupyterLab based docker images) which will be used to deploy the Jupyter notebooks. The LMSU boundaries ML pipeline will be used as its first demonstrator to be ready in spring of 2022.

The following restriction are to be noted:

1. Only registered users can run SDAs on the platform, this is needed for resource management and also to attach the user's workspace to the running SDA. In this early phase of EXPLORE the registration is upon invitation/request. In the longer-term self-registration may be added.
2. The EXPLORE platform is (currently) a development platform, which means that it has limited computing resources in the back-end. In the longer term it is foreseen to add elasticity to the infrastructure resources and evolve it to an operational service.

The deployment of tools in the ESA Datalabs is also being investigated.

A document entitled "Deployment of ML services on cloud environments", which is an update of the draft document of year 1 and which summarizes possible options for making our ML pipelines available on other platforms, was compiled.

c. Impact to date

We have a rising number of visitors on our ML Portal. At different occasions, e.g. conferences, we have presented results of our science cases as well as our ML activities in EPN-2024-RI. We have published one publication with ML contribution, there will be at least four more publications in 2022. We have organized and convened two conference sessions specifically dedicated to ML in planetary sciences and heliophysics (and we will organize such sessions again in 2022). Two workshops were conducted in the course of EPSC2021 to introduce our ML pipelines to the scientific community.

d. Summary of plans for Year 3

There is one milestone (MS51 - ML Demonstrators implemented and tested) scheduled for the end of June 2022 and one deliverable (D10.3 - Tutorial on Machine Learning and Basic How Tos (initial release)) scheduled for the end of July 2022.

At the moment, we are drafting publications with the results of the IWF ICME, the LMSU boundaries, the GMAP mounds, and the GMAP pits science cases. We will finalize the integration of first data sets of our science cases into VESPA by April 2022. Further, we will integrate first ML pipelines into SPIDER and the EXPLORE platform.

We organized two ML sessions, one at the EGU 2022 and one at the JpGU 2022. The Fireballs workshop #2, which ML organized together with NA2, will be on 4-5 February 2022; the third one in this series will be held in late fall/early winter 2022. We plan to have our next ML workshops introducing new ML pipelines in spring as well as in fall 2022.

Finally, we will start to work on the last two science cases (IAP waves, INAF spectra).

2 Update of data management plan

An update of the Data Management Plan (DMP) is due end of March 2022.

3 Follow-up of recommendations & comments from previous review(s)

The report of the VA Review Board was received in December 2020. We want to underline that WP10 is not a VA, but a JRA, and thus it cannot be reviewed in the same manner as the other VAs.
We have reacted on the comments and recommendations of the VA Review Board in the first annual report (D10.1) and have

- updated our DMP,
- set up a public GitHub account,
- looked for and investigated alternatives to EOSC,
- added more content to our ML Portal and to the GitHub repositories including tutorials for our ML pipelines, and
- set up a schedule for the integration of ML data and tools to VESPA and SPIDER and started the integration.