NOAH-H: AUTOMATIC CLASSIFICATION OF HIRISE IMAGES USING DEEP LEARNING APPLIED TO EXOMARS LANDING SITE SELECTION SUPPORT AND FUTURE MARS ROVER OPERATIONS

Virtual Conference 19–23 October 2020

Mark Woods[1], Spyros Karachalios[1], Danilo Petrocelli[1], Alexander Barrett[2], Matt Balme[2], Luc Joudrier[3]

[1]SCISYS, 23 Clothier Road, Bristol, BS45SS, UK, E-mail: mark.woods@scisys.co.uk
[2]The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK, E-mail: Alexander.Barrett@open.ac.uk
[3]ESA/ESTEC, Keplerlann 1, 2201 AZ, Noordwijk, The Netherlands, E-mail: luc.joudrier@esa.int

ABSTRACT

Manual classification of large-scale orbital images of Mars by geologists is essential to support rover landing site selection, and eventually, operations. Extensive classification is, however, often prohibitive given the scale of the data and limited availability of specialist experts. To address this challenge, we developed a Deep Learning based system called NOAH-H (Novelty or Anomaly Hunter - HiRISE) to automate this process. NOAH-H can provide pixel level, annotated classifications of terrain seen in HiRISE images, based on a 14 class ontology defined by experts in geomorphology. The system prototype was developed and evaluated during the recent ExoMars site selection process with encouraging results. This paper reports on the NOAH-H study findings.

1 INTRODUCTION

Successful rover traversability on Mars and mission viability is highly dependent on the underlying terrain type and texture of the target site. Outcrops, rugged bedrock, and dunes can all cause difficulties for rover navigation. Terrain classification with traversabilty in mind is an essential part of landing site selection and rover operations during long term route planning. This task requires specialist geological expertise to provide the core inputs, but when many sites are being considered and the area under evaluation is large, it is not always possible to provide exhaustive site analysis in the time available.

To address this challenge in the European mission context the authors conducted an investigation into using state of the art Deep Learning (DL) methods to provide automated classification of the Orbital images – specifically HiRISE [1] provided by NASA's MRO spacecraft. This investigation was carried out during ESA's ExoMars [2] landing site selection and analysis process thus providing a real application scenario. Results from the system were considered during the site analysis.

To carry out the investigation we extended previous work in the application of data driven, Deep Learning (DL) techniques for Mars terrain classification [3]. Specifically, we used a particular type of DL called Semantic Segmentation. Whilst re-using some of the toolsets developed previously, we also created an entirely new training and inference system to implement the required system. This new development was called NOAH-H.

The study consisted of three main stages namely, Ontology definition and labelling; Algorithmic evaluation and training; and Final system evaluation. Each of these stages is discussed in the sections that follow:

2 ONTOLOGY DEVELOPMENT AND LABELLING

The 25 cm/pixel HiRISE images cover areas which are 6 km wide and have a programmable distance of up to 60 km long. Terrain classification was based on a custom ontology which was defined specifically for this work. It focused on the final two candidate landing sites for ExoMars namely Oxia Planum and Mawrth Vallis.

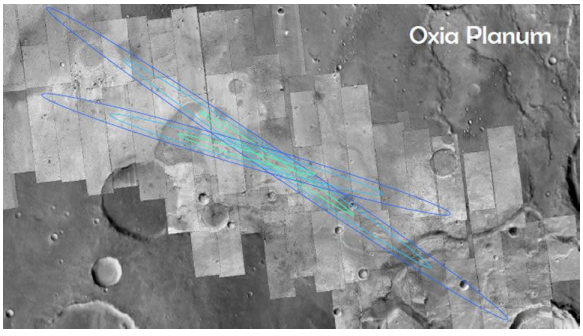

Figure 1: Oxia Planum, one of the two final ExoMars candidate landing sites (before final selection) showing landing ellipse profiles and HiRISE image coverage.

We defined a classification scheme to characterize all terrain types found within the study areas. Three broad class types with two additional sub-levels defined the hierarchical structure of the ontology.

The first level of the classification scheme consisted of three broad, interpretive groups; **Surfaces** (divided into "***bedrock***" and "***non-bedrock***" material), Aeolian bedforms or "**Ripples**" (divided into ***small*** and ***large*** features), and **Other** (which ultimately only included ***boulder fields***). We kept the interpretive elements of the scheme as basic as possible, so as to reduce the subjectivity inherent in classifying the training dataset. These groups were then subdivided into 14 ontological classes.

These classes were purely descriptive in nature, covering morphological characteristics of surface roughness and ripple morphology. They were defined purely using textural characteristics, avoiding definitions which relied upon interpreting the perceived geological origin of the material.

It was felt that the use of strict geomorphological unit labels would be impossible to implement during the labeling procedure, when contextual and situational information would be limited. Since the model would ultimately perform semantic segmentation based solely upon the morphological characteristics visible in the image, we ensured that our definitions conformed to that methodology from the start. This approach also allowed the classes to be tailored to rover traversability assessment, the primary purpose for which the model was to be used.

Distinct classes, which could be labeled with high confidence, were chosen. The classification scheme was designed to be as comprehensive as possible, so that every terrain type which the machine learning system was likely to encounter within these study areas was defined. The classification scheme of [4] was used as a starting point, and then adapted to the terrains of the candidate landing sites. These 14 classes represented the textural variations which would be most significant for traversability analysis. More detail on how the classification scheme was developed and the formal definitions of the different classes will be presented in [5].

Surfaces
Non-bedrock
1. Smooth, Featureless
2. Smooth, Lineated
3. Textured "Non-Bedrock"

Bedrock
4. Smooth "Bedrock"
5. Textured "Bedrock"
6. Rugged "Bedrock"
7. Fractured "Bedrock"

Ripples
Large Ripples
8. Simple form large ripples, Continuous
9. Simple form large ripples, Isolated
10. Rectilinear form large ripples

Small Ripples
11. Continuous small ripples
12. Non-continuous small ripples, Bedrock substrate
13. Non-continuous small ripples, Non-Bedrock substrate

Other
14. Boulder fields

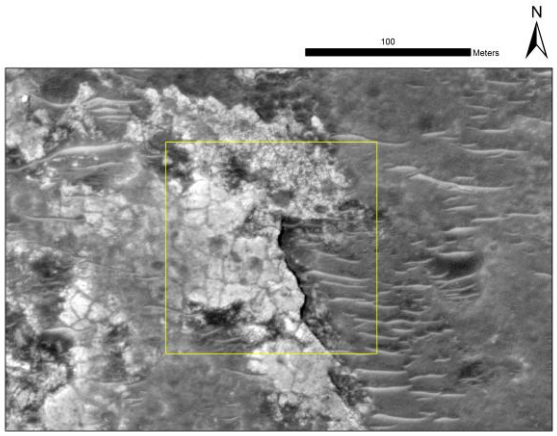

Figure 2: Example framelet. Fractured bedrock near large isolated ripples over smooth non-bedrock terrain. The yellow square is 128 m across, the size of a training framelet.

The seven surface classes define the general textures found at the site, and form a continuum of surface roughness. Terrains interpreted as "non-bedrock" have smoother overall roughness, with little surface relief. They are thus most likely to consist of unconsolidated material. Terrains interpreted as bedrock are in general rougher than the non-bedrock classes. They exhibit greater relief and sharper morphology.

Ripple classes comprise different morphologies of aeolian cover. Size is defined based upon the width of the ripple perpendicular to the ridge crest. Small features have a width of < 5m, while large features are > 5m wide. Other parameters, such as the spatial density of the ripples (whether they are continuous, discontinuous, or isolated), and whether or not they conform to a "simple sinuous" morphology are also considered. Not every combination of these descriptive parameters was included in the classification system, since not all of them occur at the site. Boul-

der fields, where discrete blocks, distinct from the underlying substrate are observed form the final class.

To provide training data for the system, the science team extracted image framelets from the full HiRISE images and used our Dataset Annotation Tool (DAT) [3], [6] to label examples of each class type. These consisted of small squares (128 m by 128 m; 512 pixels by 512 pixels). Framelets were selected to provide a suite of representative examples, from across the study areas. Each framelet contained multiple terrain types and every attempt was made to feature every observed combination of ontologies.

The DAT allowed for pixel level labeling according to the terrain types in each framelet. Areas which definitively represented each ontological class were digitized using the DAT, producing pixel-class pairs which could be used to train the model. Previous work on the NOAH system [3], where the DAT was first proposed, had relied on Citizen Science to complete the labeling. However, for this project it was decided that a large degree of expertise was required to accurately label the subtly different ontologies. The labeling work was thus conducted exclusively by the science team.

The main technical component of the labeling phase consisted of developing the DAT to allow the manual annotation and labeling of HiRISE images. The DAT was designed on top of the Oxford University Zooniverse platform [7]. This provided a lot of built-in functionality. However, certain features needed to be extended in order to better support the science team, and provide the highest quality annotation data possible. This included adding a contextual zoomed out version of the image, and providing HiRISE metadata such as image number and coordinates in latitude and longitude. The figures below show a typical example of DAT usage in the NOAH-H case with several terrain types being labeled in one Region of Interest (ROI).

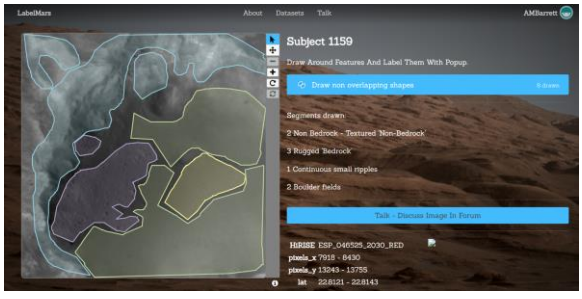

Figure 3: Example of DAT usage. (HiRISE image: ESP_046525_2030 NASA/JPL/UoA)

In total, two labeling campaigns were conducted with the second being informed by results from the first. The rationale for this is discussed further in 3.2. The labelling activity provided approximately 5,700 individual class instances across 295 MP in the final set. Approximately 66% of the available framelet pixels were labelled.

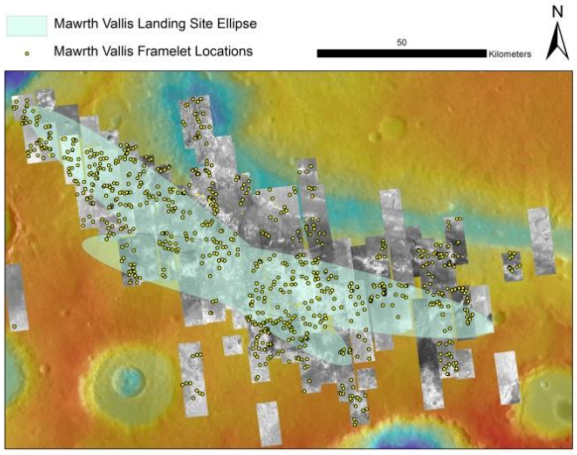

Figure 4: Map of Mawrth Vallis, showing distribution of framelets and HiRISE images after [5]. *(*[8] *NASA/JPL/UoA) over the MOLA global topographic* map [9]

3 ALGORIHTMIC APPROACH

In this work we have cast the required classification challenge as a semantic segmentation problem, allowing a system to automatically provide annotated, pixel level terrain maps, which can then be interpreted by human experts to support traversability analysis. We were not seeking to replace the human experts but augment their work, by rapidly supplying, full coverage, terrain level classifications, to inform their overall assessments. The models produce a predicted classification for every pixel in every HiRISE image under consideration, which was previously unobtainable given the limited resources available. This has the potential to greatly extend the experts' field of view and the coverage of further analysis in a time efficient way.

Given the fragmented nature of terrain distributions in the sites of interest, this was best served by providing the dense classification outputs rather than regular, larger scale geometric predictions e.g. using a bounding box assignment and object detection-based techniques. In this use case the instance assignment for contiguous terrain type groupings was considered to be part of the higher-level human assessment of

the scene where relevant. Boundary assignment is often subjective and challenging even for human experts in this type of application and often relies on much wider context.

3.1 Deep Neural Network (DNN) Methodology

Classifying whole or elements of an image has been an ongoing challenge for the computer vision community for many years. This past decade has seen significant advances being made through new developments in a specific type of learning and Neural Network based classification known as Deep Learning. Deep Neural Networks have been shown to outperform traditional or shallow neural networks and even human predictions in some image classification challenges [10].

Initially, developments and applications of DNN's have focused on image classification, object localization, and detection. As the decade has progressed, a series of progressive and iterative developments have built on the success of their predecessors and at the heart of most common architectures is a convolutional filter-based approach to feature extraction. So called Convolutional Neural Networks (CNN's) have formed the backbone of many recent developments.

This work has extended to addressing pixel level classification problems. The principle of semantic segmentation models [11–13] involves classifying each pixel of an image by analyzing the region around it, often called the receptive field, and passing it through a deep network to compute a class prediction.

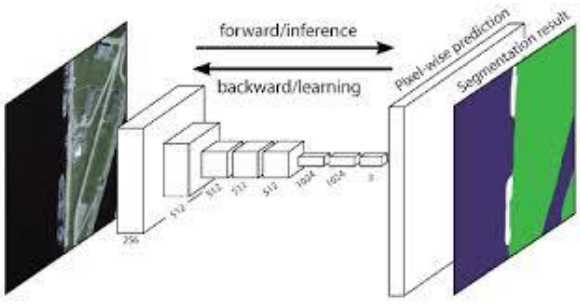

Figure 5: Fully Convolutional Neural Network Architectures for dense predictions [14].

A common state-of-the-art semantic segmentation approach has been described in [14] where the authors designed an approach to generate segmentation maps for images of any size by using a CNN architecture for dense predictions without any fully connected layers. However, some Deep Learning architectures offer a different approach to semantic segmentation by learning multi-scale contextual features. One such example is the model designed by Google: DeepLab [15–17]. Instead of regular convolutions, DeepLab uses Atrous Convolutions, also referred to as dilated convolutions, which can expand the filter's field of view.

These specialized convolutions effectively increase the receptive field of the filters without increasing the filter size. This allowed us to give more context to the network to classify each pixel while retaining the spatial relationships, which is desirable. It offered an efficient mechanism to control the field-of-view and found the best trade-off between accurate localization (small field-of-view) and wider range context with more semantic information (large field-of-view). In the course of this work we evaluated a range of architectures which realized the multiscale approach to feature extraction, incorporating bespoke modifications in order to isolate the optimum approach based on architectures available at the time of this development.

3.2 Evaluation, Training & Additional Labelling Campaigns

We measured the accuracy of our models based on standard metrics such as Precision, Recall and the Intersection over Union (IoU). All are computed using the class-wide confusion matrix which records ground truth versus prediction performance. The IoU metric is widely used for to evaluate and compare semantic segmentation models and allows us to measure the agreement between ground truth and predictions as a fraction of the total number of labelled pixels. The IoU for each class was measured using the following formula:

$$IoU = TP/(TP + FP + FN) \qquad (1)$$

Where:

- True Positive (TP): Number of pixels correctly classified.
- False Positive (FP): Number of pixels incorrectly classified as belonging to a specific class.
- False Negative (FN): Number of pixels incorrectly not classified as not belonging to a specific class.

In addition to producing an IoU for each class, a mean IoU for the combined classes is computed to provide a single number performance value. During the evaluation phase our preliminary test results and subsequent analysis revealed better than expected initial results with performance as measured by IoU in excess of 90% for some classes. Mean IoU at the full granular levels was over 70% and over 90% for the combined classes. In an attempt to improve the system performance we initiated a second labelling campaign knows as "The Second Run". This batch included new data appended to the original "First Run" dataset ensuring an overall increase in support for the various classes. A third dataset known as the "Final Run" balanced the support for some classes by making small but focused modifications to the second run. In order to properly evaluate the system, the dataset was split into two sets: Training and Validation during every run.

Statistic	First Run	Second Run	Final Run
Total Images	917	1507	1507
Total number of Images for Training	824 (~ 90%)	1414 (~ 94%)	1414 (~ 94%)
Total number of Images for Validation	93 (~ 10%)	93 (~ 6%)	93 (~ 6%)
Total number of pixels	240 MP	395 MP	395 MP
Total number of labelled pixels	151 MP	259 MP	236 MP
Percentage labelled of each image	~ 63.1%	~ 65.8%	~ 59.7%

Table 1: General dataset statistics

Table 1 shows the total number of images for each set and the total number of Pixels (given in terms of Megapixels, MP) of all images. It notes the number of MP, along with an approximation of the percentage of each image labelled using the DAT tool. To evaluate the performance of the NOAH-H system as a broader semantic segmentation network, the ontological classes were also combined into the five second order groups from the hierarchical classification scheme; Bedrock, Non-Bedrock, Large Ripples, Small Ripples, and Boulder Patches. This allowed us to assess different levels of performance as in some cases, grouped ontologies were sufficient to support the traversability analysis.

4 RESULTS

The images selected for NOAH-H classification were chosen based on criteria of low-noise, full resolution (~25 cm/pixel), and central coverage of the landing ellipse.

4.1 Output Examples

Post-processing of the NOAH-H output included down-sampling to 2m/pixel and conversion to color classes for analysis. Each of the classes was given a unique combination of RGB values in-order to create masks with the classification.

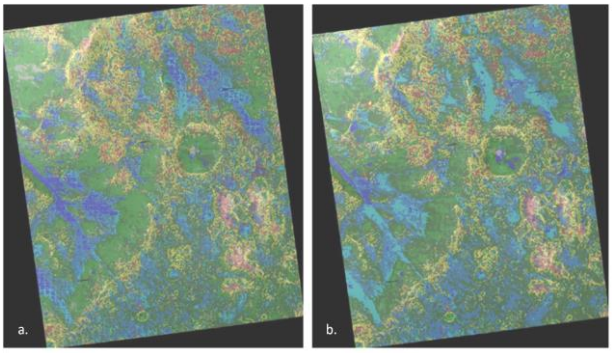

Figure 6: (a). Example Output of the "First Run" (b). Output of the "Final Run".

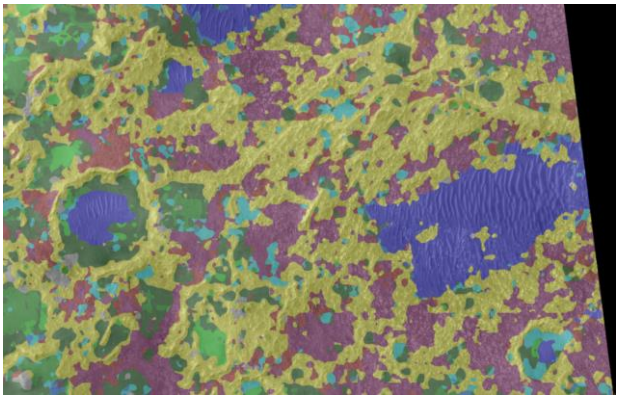

Figure 7: Close-up illustration of NOAH-H outputs.

This allowed the output to be displayed conveniently as a different color for each class in image-viewing

software, and then to be easily converted into a vector or single band product in a GIS to formally manipulate the data. GIS can also be used to overlay the NOAH-H output onto the original HiRISE images for inspection, which proved the most effective way for the science team to utilize the output raster.

4.2 Assessment

The core model was evaluated at various levels of granularity. The figures below show IoU based accuracy results for the various dataset runs.

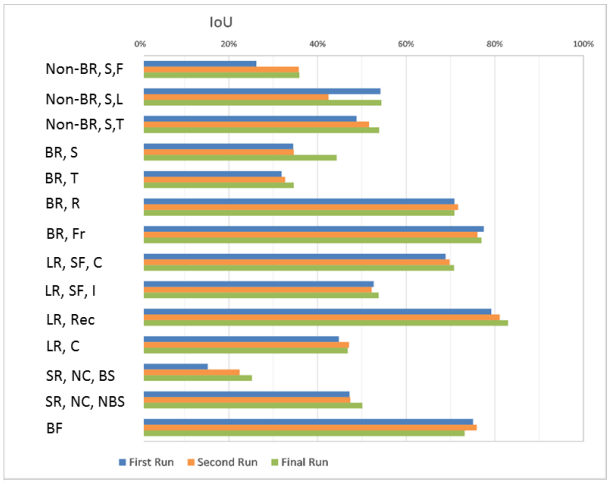

Figure 8: Accuracy results for the full 14 classes

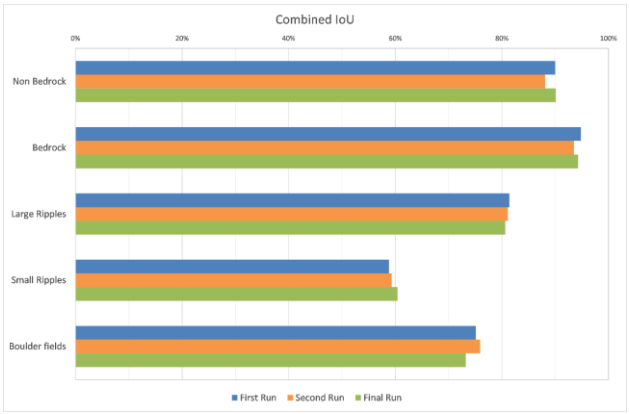

Figure 9: Accuracy results for summary class predictions.

The results demonstrate a varied level of classification performance across the 14 classes but a very encouraging overall performance, particularly given the small size of the datasets used. The system has performed very well when predicting classes such as large ripples, boulder fields, and some bedrock types. Aggregated performance for the meta-class types or groups of ontologies showed even better performance with an IoU of over 90% in some cases.

The size of the datasets does have an impact on the conclusions we draw from the work. As noted, the system provided excellent results despite relatively low levels of data by DNN standards. Our dataset is, we believe, the largest of its kind (i.e. labeled planetary remote sensing images) available at this time. It was a major achievement in itself given that it was carried out in parallel with the ExoMars site assessment with expert time being a limited resource. However, at a statistical level its relatively modest size must be borne in mind when considering an extrapolation of the application beyond the datasets used here. It is however encouraging.

In addition to the standard IoU based assessment we also carried out qualitative assessments of the system with the science team in order to better understand the system performance. The analysis highlighted several key factors which showed the strengths of the system and also areas for improvement.

The system operates best at a landscape level where contiguous groups are presented to the end user for manual evaluation and final assessment. This was in line with the original scope of the work and showed that it could provide large volumes of first order classification for expert assessment in minutes. It is not possible to replicate this type of dense prediction at scale with the limited number of experts available.

Incidences of single pixel errors in contiguous patches of terrain are a negligible issue in this context but if the outputs were to be immediately used by downstream automated processes this could cause issues. As noted in the introduction however this was not considered part of the original system scope.

Figure 10 (a) shows an interesting example of how the system can augment and speed-up manual analysis. In this case, there is a shadow area present in the crater, which would ordinarily require the use of post-processing tools in order to complete a human assessment. Figure 10(b) shows the automatic output from the NOAH-H system using the native HiRISE images which is available immediately.

Class imbalance is a major factor and a standard issue for many Machine Learning based applications. Several class types had poor support. In some cases, this did not reduce prediction accuracy excessively but in others this was clearly an issue.

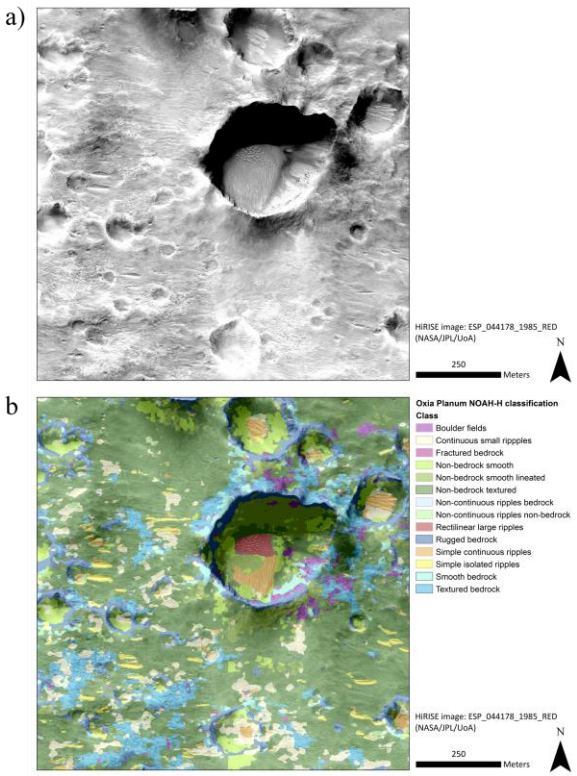

Figure 10: (a) Original HiRISE image (b)Overlaid NOAH-H Output Mask

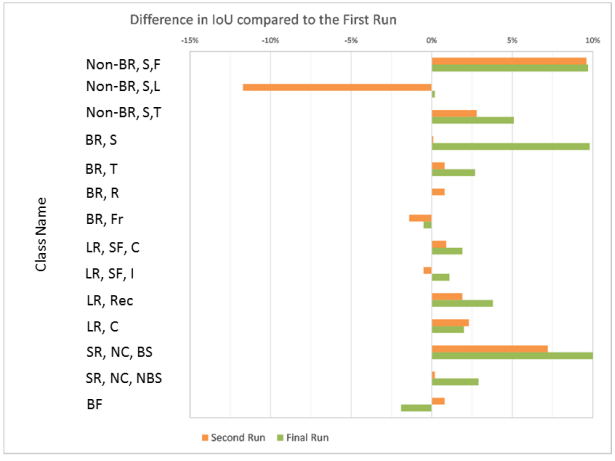

Figure 11: Accuracy deltas for Second and Final Run dataset versus the original First Run.

This was one of the drivers for our Second Run labelling campaign and re-design in the Final Run but in practice it was hard to manipulate the data to fully mitigate the under-support in some cases as the figure above shows.

This was exacerbated by other factors such as intra-class similarity and subjectivity in the original assessments. The delineation is not always clear to the human experts, even though they have the benefit of wider geomorphological data and interpretation. Broad levels of context are a crucial element of any geology based assessment and clearly its role here has been limited to a relatively localized, spatial influence on training at the pixel level. This impact was most pronounced again on class boundaries and cases where there were interactions or class "bleed" in graduated cases. Of course the reduction of the inspection task to a 14 class ontological hierarchy is the result of a trade-off between expressive richness and labelling/training practicality so inevitably there will be some confusion present in the training data which will affect performance.

5 CONCLUSION

This small study investigated the use of DNN technology to support traversability analysis for ESA's forthcoming ExoMars Rover mission. The initial results have been encouraging and the outputs from the system are currently being used to support the ongoing site investigation and analysis by the science team. The authors continue to collaborate on this and other further applications of this technology in other Space domains. This shows the value of surgically applied AI when it is used to augment human experts and carry out data intensive first order classification.

Acknowledgement

MW and SK acknowledge funding from a European Space Agency contract (ref =4000118843/16/NL/LvH–Novelty or Anomaly Hunter (NOAH). AMB and MRB acknowledge funding from the UK Science and Technology Facilities Council (STFC; grant ST/T000228/1)

References

[1] McEwan A. The High Resolution Imaging Science Experiment (HiRISE) during MRO's Primary Science Phase (PSP). *ICARUS* 2010; 205: 2–37.

[2] Vago JL, Westall F, Pasteur Instrument Teams, Landing S, et al. Habitability on Early Mars and the Search for Biosignatures with the ExoMars Rover. *Astrobiology* 2017; 17: 471–510.

[3] Read N, Woods M, Karachalios S. Novelty Or Anomaly Hunter – Driving Next Generation Science Autonomy With Large High Quality Dataset Collection. In: *I-SAIRAS 2018*. Madrid, Spain, 2018.

[4] Rothrock B, Kennedy R, Cunningham C, et al. SPOC: Deep Learning-based Terrain Classification for Mars Rover Missions. In: *AIAA SPACE 2016*. Long Beach, California: American Institute of Aeronautics and Astronautics. Epub ahead of print 13 September 2016. DOI: 10.2514/6.2016-5539.

[5] Barrett A, Balme M, Woods M, et al. NOAH-H, a deep-learning, terrain analysis system: results for the ExoMars Rover candidate landing sites. in preparation.

[6] Schwenzer SP, Woods M, Karachalios S. LABELMARS: CREATING AN EXTREMELY LARGE MARTIAN IMAGE DATASET THROUGH MACHINE LEARNING. In: *Lunar and Planetary Science XLVIII 2017*. Houston, tX, USA, 2017, pp. 1–2.

[7] Simpson R, De Roure D. Zooniverse: observing the world's largest citizen science platform. In: *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. 2014, pp. 1049–1054.

[8] McEwen AS, Banks ME, Baugh N, et al. The High Resolution Imaging Science Experiment (HiRISE) during MRO's Primary Science Phase (PSP). *Icarus* 2010; 205: 2–37.

[9] Smith E, Zuber MT, Frey V, et al. Mars Orbiter Laser Altimeter : Experiment summary after the first year of global mapping of Mars. *Journal of Geophysical Research* 2001; 106: 689–722.

[10] He K, Zhang X, Ren S, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv:150201852 [cs]*, http://arxiv.org/abs/1502.01852 (2015, accessed 11 September 2020).

[11] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2017; 39: 2481–2495.

[12] Chen L, Zhu Y, Papandreou G, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv e-prints* 2018; arXiv:1802.02611.

[13] Papandreou G, Chen L, Murphy KP, et al. Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation. In: *The IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1742–1750.

[14] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440.

[15] Chen L. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv e-prints* 2017; 1706.05587.

[16] Chen L, Papandreou G, Member S, et al. DeepLab : Semantic Image Segmentation with Deep Convolutional Nets , Atrous Convolution , and Fully Connected CRFs. *IEEE Transactions on Pattern Recognmition and Machine Intelligence* 2018; 40: 834–848.

[17] Liu C, Chen L, Schroff F, et al. Auto-Deeplab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 82–92.