

Open Research Online

The Open University's repository of research publications and other research outputs

NOAH-H, a deep-learning, terrain classification system for Mars: Results for the ExoMars Rover candidate landing sites

Journal Item

How to cite:

Barrett, Alexander M.; Balme, Matthew R.; Woods, Mark; Karachalios, Spyros; Petrocelli, Danilo; Joudrier, Luc and Sefton-Nash, Elliot (2022). NOAH-H, a deep-learning, terrain classification system for Mars: Results for the ExoMars Rover candidate landing sites. *Icarus*, 371, article no. 114701.

For guidance on citations see [FAQs](#).

© 2021 The Authors.



<https://creativecommons.org/licenses/by-nc-nd/4.0/>

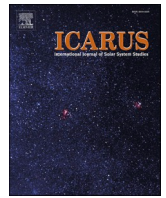
Version: Version of Record

Link(s) to article on publisher's website:

<http://dx.doi.org/doi:10.1016/j.icarus.2021.114701>

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk



NOAH-H, a deep-learning, terrain classification system for Mars: Results for the ExoMars Rover candidate landing sites

Alexander M. Barrett^{a,*}, Matthew R. Balme^a, Mark Woods^b, Spyros Karachalios^b, Danilo Petrocelli^b, Luc Joudrier^c, Elliot Sefton-Nash^c

^a Open University, Walton Hall, Milton Keynes MK7 6AA, UK

^b SCISYS Ltd, Methuen Park, Chippenham SN14 0GB, UK

^c ESTEC, European Space Agency, Noordwijk, the Netherlands

ARTICLE INFO

Keywords:

Machine learning
Mars surface
Geomorphology
ExoMars

ABSTRACT

In this investigation a deep learning terrain classification system, the “Novelty or Anomaly Hunter – HiRISE” (NOAH-H), was used to classify High Resolution Imaging Science Experiment (HiRISE) images of Oxia Planum and Mawrth Vallis. A set of ontological classes was developed that covered the variety of surface textures and aeolian bedforms present at both sites. Labelled type-examples of these classes were used to train a Deep Neural Network (DNN) to perform semantic segmentation in order to identify these classes in further HiRISE images.

This contribution discusses the methods and results of the study from a geomorphologists perspective, providing a case study applying machine learning to a landscape classification task. Our aim is to highlight considerations about how to compile training datasets, select ontological classes, and understand what such systems can and cannot do. We highlight issues that arise when adapting a traditional planetary mapping workflow to the production of training data. We discuss both the pixel scale accuracy of the model, and how qualitative factors can influence the reliability and usability of the output.

We conclude that “landscape level” reliability is critical for the use of the output raster by humans. The output can often be more useful than pixel scale accuracy statistics would suggest, however the product must be treated with caution, and not considered a final arbiter of geological origin. A good understanding of how and why the model classifies different landscape features is vital to interpreting it reliably. When used appropriately the classified raster provides a good indication of the prevalence and distribution of different terrain types, and informs our understanding of the study areas. We thus conclude that it is fit for purpose, and suitable for use in further work.

1. Introduction

The ExoMars *Rosalind Franklin* rover and *Kazachok* surface platform (Vago et al., 2017), is expected to land at Oxia Planum in 2023. This mission will search for signs of past and present life. This paper describes NOAH-H (Novelty or Anomaly Hunter - HiRISE) an ESA funded project conducted in 2018 as a collaboration between the Open University, the ExoMars Landing Site Selection Working Group (LSSWG), and the SCISYS Autonomy & Robotics Group.

This consisted of an investigation into automatic terrain classification using Deep Learning (DL). The aim was to automatically identify metre to decametre-scale terrain types using 25 cm/pixel High Resolution Imaging Science Experiment (HiRISE) images (McEwen et al.,

2010). The study area consisted of the final two ExoMars candidate landing sites: Oxia Planum and Mawrth Vallis (Loizeau et al., 2019). The training dataset was tailored to the characteristics of those sites. The task of terrain classification was framed as a semantic segmentation problem. The aim was for the Deep Neural Network (DNN) to classify each pixel in a HiRISE image according to a prescribed ontological class, and so produce surface texture terrain maps for the sites.

This paper provides an overview of the methods of the NOAH-H project, and an evaluation of its results. The focus is on the definition of the classification scheme and geomorphological analysis of the output raster. We examine the machine learning workflow from a geomorphologist's perspective, highlighting issues to consider when developing classification schemes and applying them to often complex

* Corresponding author.

E-mail address: alexander.barrett@open.ac.uk (A.M. Barrett).

<https://doi.org/10.1016/j.icarus.2021.114701>

Received 30 July 2020; Received in revised form 3 September 2021; Accepted 10 September 2021

Available online 13 September 2021

0019-1035/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

landscapes. We demonstrate that, when considered at a landscape level, the reliability and fitness for purpose of the model must be considered qualitatively as well as quantitatively.

The model output is intended to provide one component of rover traversability assessment, when combined with information about the topography of the site, and rover engineering parameters. A detailed examination of traversability is important for landing site selection, pre-mission planning, and ongoing rover operation. Terrains such as large bedforms or outcrops of rugged bedrock may complicate rover navigation or prevent it from reaching its science targets. Regions of unconsolidated non-bedrock can cause rovers to become stuck. However, because of the large size of the potential landing ellipse, manually identifying hazardous areas at HiRISE scale is prohibitively time consuming. This necessitates a machine learning approach. The issue of traversability is used as an example throughout this paper, however a full hazard analysis is not presented as this is beyond the scope of the current contribution.

Section 1 introduces the background to the project, and states the case for a machine learning approach. In Sections 2 and 3 the ontological classes are introduced, and the process for developing them explained. Section 4 describes the acquisition of the training dataset. Section 5 presents results for pixel-scale accuracy. Section 6 compares these to a qualitative evaluation of representative areas of classified terrains. It assesses the reliability of the model at a landscape, rather than pixel to pixel scale. These results are further examined in sections 7 and 8, where considerations and limitations are discussed.

1.1. Background

The ever-increasing volume of data being returned by planetary remote sensing missions, and its increasingly high spatial resolution, presents both a great opportunity for the science of planetary geomorphology, and a major challenge. The surfaces of other planetary bodies can be studied in unprecedented detail. However, the more high-resolution data is acquired, the less amenable to complete study the dataset becomes. The time required to survey data products at full resolution means that comprehensive studies are becoming impractical for all but the largest teams. Machine Learning (ML) presents a powerful tool to overcome these challenges by automating certain aspects of dataset interrogation. This method is thus valuable, not just for the ExoMars mission, but for all future efforts in planetary exploration (e.g. Harris and Martin, 2020). The application of ML to planetary remote sensing (RS) data is a fast growing field.

Mars is a popular target for machine learning studies due to the large amount of RS data available. Studies using multispectral data such as that from the Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) have seen success in identifying mineral assemblages which would have been challenging to isolate by conventional means. (e.g. Dundar and Ehlmann, 2016; Saranathan and Parente, 2021). The scale and depth of multispectral data means that the development of automated systems to analyse the data in a timely fashion has long been of high priority (e.g. Allender and Stepinski, 2017; Lin et al., 2018; Parente et al., 2011).

The same is increasingly true for studies focused on geomorphology, which generally employ visible band images, topographic data, or some combination thereof. The most common objective for ML in planetary geomorphology is the detection of craters (e.g. Bandeira et al., 2012; Cadogan, 2020; Silburt et al., 2019; Stepinski et al., 2009; Urbach and Stepinski, 2009; Wang and Wu, 2019). Calculating crater density remains the primary method for dating planetary surfaces (Hartmann and Neukum, 2001). A tool to count craters automatically would thus be very useful (e.g. Salamunicar et al., 2012 and references therein).

There has been a shift in focus in the past 4–5 years (Wilhelm et al., 2020). Early geomorphology ML studies (e.g. Ghosh et al., 2010; Jasiewicz and Stepinski, 2012; Stepinski and Vilalta, 2005) are sparse, and primarily employed low resolution topographic data from MOLA

(Mars Orbiter Laser Altimeter). For example Jasiewicz and Stepinski (2012) proposed a comprehensive terrain classification system based on a “geomorphometric map”. They used MOLA to identify generic variations in topography. This served as an input for the automatic detection of geomorphological units which exhibit the same topographic signature, independent of scale. Work using visible band images was less common and limited to the Mars Orbiter Camera (MOC) (e.g. Bandeira et al., 2013, 2011) even in the years after higher resolution images became available.

However, since 2016 ML studies have become more frequent, and many have used the high resolution cameras on Mars Reconnaissance Orbiter; HiRISE (e.g. Foroutan and Zimbelman, 2017; Palafox et al., 2017; Rothrock et al., 2016a; Wang et al., 2017), and the Context Camera (CTX) (e.g. Palafox et al., 2017; Wilhelm et al., 2020). Wilhelm et al. (2020) note that the shift to high resolution visible images coincides with the adoption of Deep Learning (DL) techniques (LeCun et al., 2015). The advent of DL likely resulted in an increase in interest in applying ML techniques to the ever expanding HiRISE catalogue (Wilhelm et al., 2020).

By necessity, the majority of studies only attempt to segment a limited suite of features. Most focus on one or two thematic groups and develop models to distinguish these from a background. Pina et al. (2008) have mapped the distribution of polygons. Foroutan and Zimbelman (2017) identified bedforms using self-organising maps, while Bandeira et al. (2011) focused on dunes.

More general studies include Palafox et al. (2017) who identified a variety of cones and bedforms. Ghosh et al. (2010) and Wilhelm et al. (2020) mapped more comprehensive suites of landforms. Rothrock et al. (2016) is the closest precursor to the present work. They investigated the application of DL techniques to rover traversability in both in situ images and HiRISE data. Other studies to examine rover traversability using in situ data include Harris and Martin (2020); and Karachalios et al. (2019).

1.2. NOAH-H

NOAH-H is the latest stage in a series of research activities carried out by the SCISYS Autonomy & Robotics Group in applying machine learning and computer vision techniques to robotic planetary exploration (Schwenzer et al., 2019; Wallace et al., 2017; Wallace and Woods, 2015; Woods et al., 2015, 2009).

As DL based approaches to image understanding have advanced in recent years (LeCun et al., 2015), we carried out an early investigation of their applicability to Mars rover based scene understanding (Karachalios et al., 2019). Such algorithms learn relationships based on labelled examples or training data. To address this challenge, dedicated planetary toolsets were developed. Crowd sourced labelling campaigns were undertaken to label large numbers of Mars landscape images acquired during rover navigation (Schwenzer et al., 2019; Wallace et al., 2017).

As the ESA ExoMars mission evolved there was interest in exploring whether DL techniques could be extended to classify images from Mars orbit, in support of landing site selection. The original NOAH (Novelty or Anomaly Hunter) project was therefore extended to classify large volumes of HiRISE images, by applying advanced DL techniques to the problem of pixel-level terrain classification in orbital images.

The purpose of this tool is not to replace the human geomorphologist, and produce a perfect geomorphological map. Rather it is intended to augment a geomorphologist’s workflow, by essentially performing “triage” on the vast catalogue of HiRISE data. The model highlights regions of interest, and gives the human operator information about the distribution and prevalence of different terrains. This can then speed up formal geomorphological mapping tasks, and provide a useful component in addressing other questions such as rover hazard assessment. However, it is not intended to be a sole arbiter of either the geological origin of the terrain, or the hazard it might pose to a rover. Rather it adds value to the original image, providing the human expert with a new tool

to inform their understanding of the terrain.

2. Methods

Geomorphic classes were defined by the science team, and manually labelled in small “framelets” (128 m by 128 m; 512 pixels by 512 pixels) extracted from red-band HiRISE images. These provided representative coverage of the study areas. Framelet locations were chosen specifically to include multiple terrain types, in various combinations.

Many prior studies into ML terrain classification have relied upon digital elevation data as well as or instead of visible band images (e.g. [Bandeira et al., 2013, 2011](#); [Jasiewicz and Stepinski, 2012](#)). This has various advantages, especially for tasks such as crater detection where the change in elevation is characteristic of the target feature.

However, we decided not to employ elevation data. While there is global coverage of low resolution data from MOLA, high resolution topographic data remains rare. Digital elevations models (DEM) at a comparable scale to HiRISE images are generally produced by combining stereo HiRISE images using photogrammetry. Thus only a fraction of the HiRISE catalogue can be used to produce DEMs, and they have small enough coverage that a machine learning approach to classification is largely unnecessary. While our site has relatively good stereo coverage, it was not practical to make DEMs for every available image pair in the area. Consulting multiple datasets would also have complicated the procedure, making the dataset more laborious to label reliably, and more complex for the ML algorithm to interpret.

Instead it was decided to use only the original red-band HiRISE images. Using colour HiRISE images was considered, however these are much narrower than the equivalent red-band image, and even in a well imaged landing site there is not full coverage. Creating an algorithm which could have used colour data in some places, but coped without it in others would also have added additional complexity and so was out of scope for the project. Consequently the high coverage red-band images were chosen.

These are important considerations when designing a project of this sort. While using additional sources of data could well be advantageous, it is important to consider whether they are strictly necessary. The need for thousands of training framelets can quickly multiply the work required to source and cross reference additional datasets.

2.1. Defining ontological classes

We produced a comprehensive classification scheme consisting of fourteen classes in five thematic groups as outlined in [Section 3](#). These classes must cover every terrain which might be encountered in the study area. The list could not be prohibitively long, as this would both complicate the training process and slow down the expert labelling work required to compile the training dataset. It was also important that classes be sufficiently distinct that they could be distinguished by the geomorphology team with a high level of confidence with limited context information. Previous studies into classification for rover traversability (e.g., [Rothrock et al., 2016](#)) informed the definitions. These were then tailored to the candidate landing sites. A wide variety of terrains are found across the martian surface, however not all are represented in the study areas. The classification system is thus specific to this and similar regions.

Classes were defined based upon the textural characteristics of the surface rather than perceived geological origin or strict geomorphological unit definitions. They describe the fundamental textures which, in various combinations, form those units. This made them general enough to be applied to multiple future science questions, giving the model maximum utility. This also allowed the classes to be descriptive, rather than interpretive.

Geomorphological interpretation requires substantial contextual and situational evidence. This would be lacking for the small snapshots used for the labelling exercise. The labelling of examples had to remain as

consistent as possible. However, in the absence of ground truth, it is impossible to be certain that a given interpretation of the landscape is correct. This approach ensured that all examples of a given texture definitively represent it, even if their final interpretation could be debated. It also reduced the subjectivity inherent in classification, and thus increased the fidelity and reproducibility of the labelling. This approach did not limit the applicability of the model, since the descriptive classes could easily be combined into broader “interpretive groups”.

2.2. Traversability considerations

For each class we suggest a likely level of hazard. This is not intended to be definitive. Rover route planning is a complex process, involving assessment of slope, hazard avoidance, and modelling of the interaction of a specific rover’s wheels and weight with soils of different compositions. This requires the input of a variety of different lines of evidence, in particular in situ observations of soil type which cannot be determined with certainty from orbit.

Rather our classification is intended to give a general indication of areas where potentially hazardous terrains are more prevalent. The model identifies areas with a high proportion of extensive localised hazards (e.g. large aeolian ripples, dense boulder fields, and fractured bedrock terrain). It also provides an overall assessment of qualitative surface roughness, showing regions which have extensive rugged terrain, and those where smoother bedrock is found. It identifies areas of non-bedrock terrain, but cannot distinguish areas of smooth regolith, which might be safe to traverse, from sandy areas which might not. This discrimination would require further analysis by a human geomorphologist, using the classified raster as a starting point and considering various lines of situational or contextual evidence such as the prevalence of aeolian bedforms coterminous with areas of smooth non bedrock terrain. Our classes can indicate which landing sites have the highest proportion of potentially hazardous terrains, and provide a useful starting point for more in depth route planning operations.

2.3. Descriptive parameters for terrain classes

A set of descriptive parameters was developed (see [Table 1](#)), which can be combined to comprehensively characterise any terrain. These consist of a series of fundamental textures, which can then be classified based on various parameters: scale, slope, pattern, distribution, and apparent substrate. Thus, while the final set of classes must be tailored specifically to a certain region of Mars, with only limited transferability to other sites, this scheme provides a robust tool for creating compatible classification schemes for future work.

These parameters can be combined into thousands of possible permutations, however not all are present at the sites, or in fact physically possible. Invalid combinations were ignored, and further surveying was conducted to determine which combinations were useful. The pattern and distribution parameters are only applicable to non-surface textures. In principle they could be applied to ripples, dunes, and clastic patterns. However, in practice, only ripple-forms were examined: dune-forms were not found in the two study areas, and boulder fields were all essentially randomly distributed, and so could be described by a single class.

It was decided that slope should not be considered as a parameter, since it could not be reliably determined, based solely on the information within a framelet. It was decided that it was more efficient to compare the classified raster to a digital elevation model downstream, than to try to train the AI to estimate slope from geomorphological indicators or include DEMs in the training dataset. Scale was only applied to the Aeolian features, since it was not applicable to surface textures, which can cover an area of any size.

The result was 14 ontological classes which cover the full variety of terrains at the landing sites, while being manageable for use in the

Table 1
Descriptive parameters.

Basic Categories	
Fundamental	Interpretation
textures	
Smooth	Smooth regolith, or flat-lying aeolian materials
Textured	Bedrock or rougher regolith surface
Rough	Rugged bedrock & outcrops
Bedform	Transverse Aeolian Ridges (TARs; e.g. Balme et al., 2008) or smaller aeolian ripple-like forms.
Duneform	Dunes
Fractured	Probably occurring in bedrock
Clastic	Blockfields and boulder patterns
Intrinsic Properties	
Scale	
Large (100 m)	Fills the 128 × 128 m framelet
Medium (10 m)	Substantial size in framelet, easily digitised
Small (m)	Small or subtle features. Often too small to digitise individually
Slope	
Flat	Expanse of level terrain
Slope	Slope clearly evident from morphology/shadow (e.g. crater wall)
Crest	Ridges and crater rims
Descriptive Classes for	Discontinuous Textures
Pattern	Interpretation
Irregular	No discernible pattern
Linear	Parallel features
Polygonal	Polygonal and rectilinear patterns
Distribution	Interpretation
Continuous	Total cover by a certain texture (e.g., aeolian bedforms)
Dense	Covers most of the surface area
Sparse	Covers less than half of the surface area
Isolated	Single features, separated from others
Substrate	Interpretation
Bedrock	Solid surface
Non- Bedrock	Surface consisting of regolith or aeolian drift.

labelling exercise. These are organised into a hierarchical classification scheme. The entire scheme is split into Surfaces, and types of “cover”. The 14 descriptive classes are arranged into five “interpretive groups” consisting of bedrock and non bedrock surfaces, aeolian cover on two distinct scales, and a single class for clastic cover.

This division is informed by surface roughness, and whether the class comprises a surface texture or a type of cover. For surfaces this requires distinguishing between bedrock and non-bedrock. This is a basic interpretation, which human planetary geologists make as a fundamental step in classifying a surface, using clues such as the presence or absence of subtle fracturing, or hard linear edges to the surface texture. Whether a surface comprises bedrock is vital for rover traversability – as bedrock always provides “grip” irrespective of whether it is rugged or smooth. Conversely, non-bedrock could consist of cloddy regolith, or be comprised of sand or dust, which would perhaps be impassable. These materials cannot be distinguished from orbit, only in-situ.

When used in combination the 2nd and 3rd levels of the classification scheme allow distinct textures to be labelled with high repeatability using the full class list, and broad interpretations of the landscape to be achieved using the combined list. Neither ontology, in and of itself, is sufficient for traversability assessment, but when combined they can provide a lot of information about the character of a site.

3. Overview of ontological classes

Surface classes define the seven broad textures found across the site. Rough textures with sharp surface morphology can generally be interpreted as bedrock, whereas smoother areas are more likely to consist of unconsolidated material such as regolith or aeolian deposits (i.e. sand or possibly dusty surface). These are collectively referred to as non-bedrock surfaces.

The “dispersed” classes primarily describe a variety of types of

aeolian bedforms and clastic cover. These are described in terms of several parameters: ‘Size’ defined by the distance measured across the bedform, perpendicular to the ridge crest, divides all features into either “small” (<5 m) or “large” (>5–20 m). These groupings are then subdivided according to spatial density (continuous vs. isolated), and morphology (simple/sinuuous vs. rectilinear). Finally, whether they overlie bedrock, or non-bedrock material is relevant for non-continuous features, as this will affect traversability (a rover might navigate around discontinuous ripple-cover, but not a continuous field). This is the only element of interpretation which could not be avoided in the descriptive layer of the scheme. These descriptive classes form a large number of permutations, however not all are present at the site and not all are ‘allowed combinations’. Six ripple classes were chosen, as summarised below.

The final dispersed class; Boulder fields, consists of patches of discrete blocks distinct from the underlying substrate.

Each class will have slightly different implications for traversability, which are discussed in the full description below. Table two lists which are expected to have high, low, or very low traversability, as well as classes for which the progress of the rover will be controlled by local hazards, and those for which the traversability is uncertain without in situ observations. Examples of these features are shown in Figs. 1–3. (See Table 2).

3.1. Non-bedrock surfaces

These can be interpreted as either regolith or aeolian surfaces, depending on specific morphology.

3.1.1. Class 1: smooth, featureless

These areas have a smooth surface in HiRISE images, and low relief at metre and 5–10 m scale. There is no evidence for bedrock. The traversability of these terrains is uncertain, since their exact composition cannot be established from orbital images. As a traversability class they should be treated as potentially hazardous.

3.1.2. Class 2: smooth, lineated

Slope mantling materials can be distinguished from flat regolith surfaces by subtle lineations influenced by the direction of slope. They have few surface markings and few rocks visible in HiRISE images. They are usually associated with steep slopes such as crater walls. This means that traversability will be very poor, since the steep slope, and potentially unconsolidated material would be hazardous for the rover.

3.1.3. Class 3: textured non-bedrock

This class is similar to “Smooth, Featureless” but with some minor texture, smooth at the 5–10 m scale. There is no evidence for bedrock. Since the material is unconsolidated it could pose a hazard, but this could only really be determined in situ. Caution should thus be taken when making an assessment from remote sensing data.

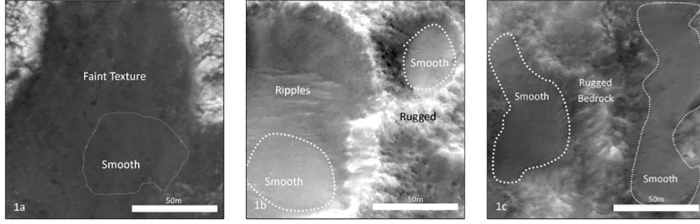
3.2. Bedrock surfaces

These classes comprise generally rougher terrains indicative of exposed bedrock. This in turn implies competent surface rock, so traversability will mainly depend upon the metre-scale relief.

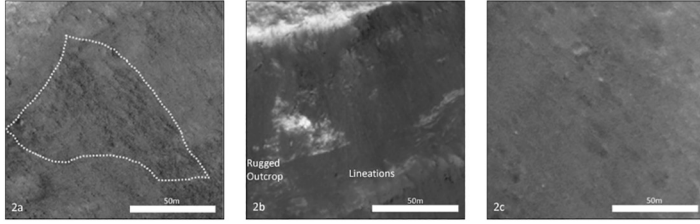
3.2.1. Class 4: smooth, “bedrock”

This class consists of patches of smooth, flat-lying bedrock. They are often bright when compared to mantling materials. They have enough texture to distinguish them from the smoother non-bedrock terrains, but are not rough enough to be classified as textured or rugged bedrock. Smooth bedrock would likely have good traversability. Bedrock provides grip for the rover, and the smooth surfaces present relatively few large localised hazards.

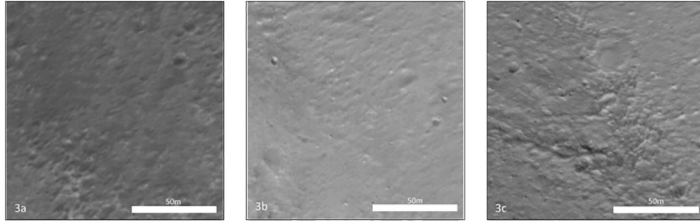
Class 1: Smooth, featureless non-bedrock



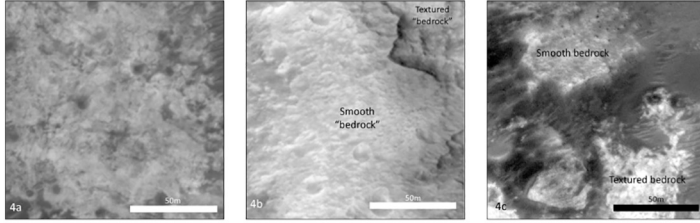
Class 2: Smooth, lineated non-bedrock



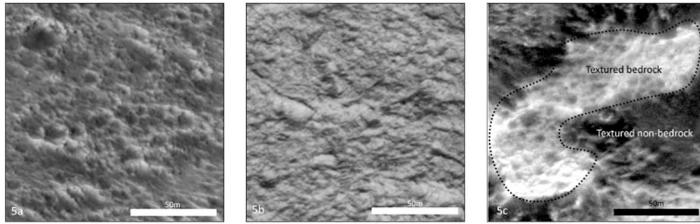
Class 3: Textured non-bedrock



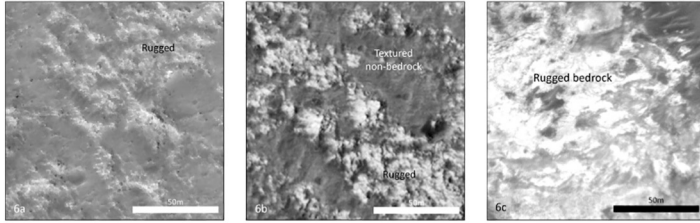
Class 4: Smooth bedrock



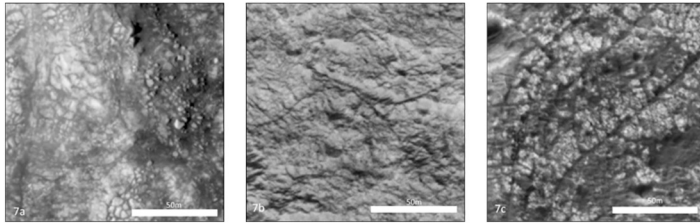
Class 5: Textured bedrock



Class 6: Rugged bedrock

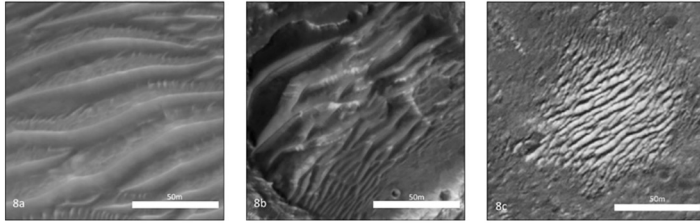


Class 7: Fractured bedrock

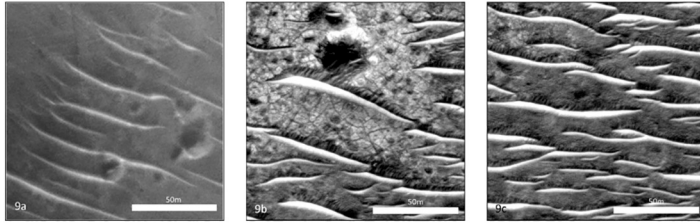
**Fig. 1.** Surface texture class examples.

Three examples of each class are shown, demonstrating the variety of forms included. Each image is 128–128 m, the size of a training framelet. The scale bar is 50 m, and north is up. HiRISE image codes can be found at the end of the respective captions. All HiRISE products shown or referenced throughout this paper are credited to NASA/JPL/University of Arizona. 1. Smooth featureless. Coherent, textureless material grades to textured non-bedrock at the edges (1a). Smooth featureless material often pools in topographic lows, such as small impact craters (1b), or between rugged outcrops (1c). (ESP_045457_2025, ESP_042134_1985, ESP_036661_2025). 2. Smooth lineated. Generally found on crater walls (2b&c). Lineations occur parallel to the direction of slope. Slopes are generally smooth but with slight streaks. Lineated terrains can be pronounced (2a) or subtle (2b&c). (ESP_046960_2030, ESP_040288_1980, ESP_045523_1985). 3. Textured non-bedrock. Generally smooth patches of non-bedrock material with clear pits or undulations on the <5 m scale. Often found in conjunction with featureless material, as one type is frequently found to grade into another (3b). Texture in 3b is less pronounced than in 3a it grades into lineated terrain to the south, and possible ripple-like forms to the north. 3c exhibits the most texture, with clear bumps and depressions. (ESP_036661_2025, ESP_037703_1980, ESP_044204_2020). 4. Smooth bedrock. Often consists of bright material with little surface roughness. Small patches of “smooth featureless” terrain can form within low points on the lightly textured surface. These are particularly evident at the sides of 4a. The small blocks in 4b and 4c have a few pits but are not as rough as the textured bedrock also seen in 4c. Only small areas of this type are typically found within the study areas, they often occur in close proximity to textured bedrock (4c), or grade into it (4a&b). (ESP_045114_2025, ESP_046156_1980, ESP_036661_2025). 5. Textured bedrock. Consists of areas of rougher textured rock. The ground is often pitted by many small craters (5a). Or areas of bedrock can exhibit small undulations, furrows, and ridges on a 5–20 m scale (5b). These often form blocks surrounded by non-bedrock material (5c). (ESP_033826_2030, ESP_042556_1985, ESP_051351_2025). 6. Rugged bedrock. The roughest bedrock surfaces, with the most pronounced texture and the highest relief. In 6a bright areas consist of patches of rugged bedrock emerging from beneath the darker, smooth featureless, mantling material. Small areas of ripples are seen around the bedrock outcrop. Bright outcrops of rugged bedrock can be interspersed with darker patches of smooth featureless and textured non-bedrock terrain (6b), or grade into textured or smooth bedrock (6c). (PSP_007019_1980, ESP_037070_1985, ESP_036661_2025). 7. Fractured bedrock. Areas of bright bedrock, clearly fractured in a polygonal or rectilinear pattern. In 7a this fracturing becomes more pronounced, and darker to the east, where smooth material can be seen in the gaps between fractured blocks. In 7b Small networks of fractures occupy the centre, incising the textured bedrock which covers the rest of the framelet. These fractures are long, sinuous features, with only a few polygonal cells. In 7c a much more complex network of fractures incises rugged bedrock. This network covers a larger area and has clear polygonal and rectilinear cells. Some patches of smooth textured terrain can be seen in the darker regions between fractures and blocks. (ESP_045747_2030, ESP_042556_1985, ESP_044679_1985).

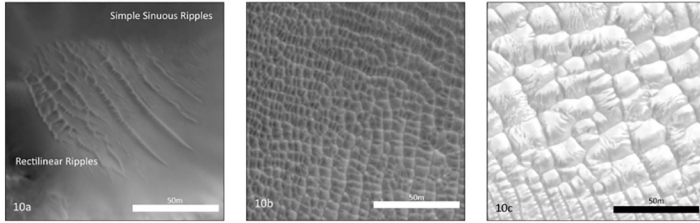
Class 8: Simple form large ripples, continuous



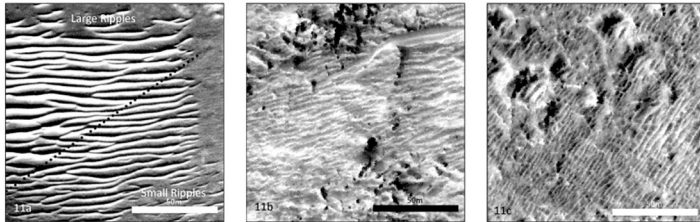
Class 9: Simple form large ripples, isolated



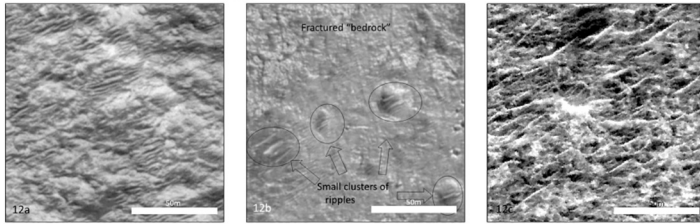
Class 10: Rectilinear form large ripples



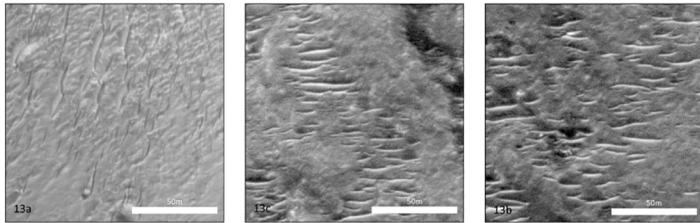
Class 11: Continuous small ripples



Class 12: Non-continuous small ripples, bedrock substrate



Class 13: Non-continuous small ripples, non-bedrock substrate



Class 14: Boulder fields

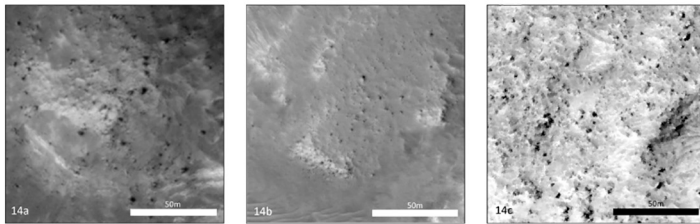


Fig. 2. Aeolian and clastic cover class examples. 8. Simple form large ripples, continuous. Continuous fields of decametre scale ripples. All of the material between the ridge crests has the same texture as the ripples themselves, so can be interpreted to be an aeolian deposit (e.g. 8a). 8b shows variation in size within this class with transitional ripples in a small crater. Features to the top left are very large but become progressively smaller towards the bottom left, before dropping below the 5 m threshold for “large” ripples. Features in 8c have the same morphology, but are on the smallest end of the scale. Most are only slightly larger than the < 5 m cut-off. (ESP_036661_2025, ESP_037070_1985, ESP_011937_1970). 9. Simple form large ripples, isolated. Isolated large ripples over any substrate. The ground between the ripples can have a smooth featureless texture (9a), textured (9c), or bedrock (9b). The surface in 9a could be of the same material which makes up the bedforms, however this cannot be definitively determined. The ridge crests are separate, and the structures do not merge together, except at the ends of a few ripples. (ESP_045523_1985, ESP_046525_2030, ESP_046525_2030). 10. Rectilinear form large ripples. A rare class, where perpendicular banks of ripples intersect forming a network of rectangular cells. 10a shows the transition between rectilinear and simple forms; rectilinear cells can be seen to the left side of the image and grade into longer, sinuous ripples with a simple morphology or parallel ridge crests. 10b shows medium sized bedforms while 10c shows larger ones. All exhibit a clear rectilinear pattern. In these cases cells cover the entire area of the framelet. (ESP_036872_2025, ESP_040288_1980, ESP_044204_2020). 11. Continuous small ripples. Small ripples < 5 m across, which form a continuous blanket, with no intervening material. These are often found on the periphery of patches of large continuous ripples (11a), however, only those ripples which are entirely below 5 m in width qualify. Patches of tiny ripples cover large areas, some are only just above the resolution of the image (11b&c). (ESP_049162_2020, ESP_036925_1985, ESP_037070_1985). 12. Non-continuous small ripples, bedrock. Small < 5 m ripples which are sparsely distributed over bedrock substrates. In 12a dense, but non-continuous, bedforms, form patches which overlay textured bedrock. In 12b. Very small patches of discontinuous bedforms are found within small impact craters. The majority of the image is made up of textured and fractured bedrock 12c shows slightly larger bedforms, which are more evenly spread over the area of bedrock. (ESP_042556_1985, PSP_002694_1985, ESP_040433_1985). 13. Non-continuous small ripples, non-bedrock. Small < 5 m ripples which are sparsely distributed over non-bedrock substrates. 13a shows Approximately evenly spaced non-continuous bedforms, trending north-south. The material between them has little texture, suggesting that it is of a non-bedrock type. 13b&c show larger bedforms, on the upper limit of the size class. Some patches of bedrock protrude from the smooth featureless mantle, but the majority of ripples occur over non bedrock. (ESP_043637_2030, ESP_044204_2020, ESP_044204_2021). 14. Boulder fields. Areas with denser boulder cover. In 14a the clastic material over-lies a region of textured bedrock, and is surrounded by smooth and textured bedrock. There is space between individual clasts. Clast size varies, however the boulder field is spread fairly evenly over the area it covers. The boulder field in 14b is much less dense, with a small scattering of clasts across almost the entire area of textured bedrock, which occupies most of the framelet. Small patches of textured non-bedrock can be seen around the edges, and do not include clastic material. 14c shows a slightly denser, and more widespread boulder patch, in the vicinity of a small impact crater. (ESP_046960_2030, ESP_406103_2030, ESP_044811_1985).

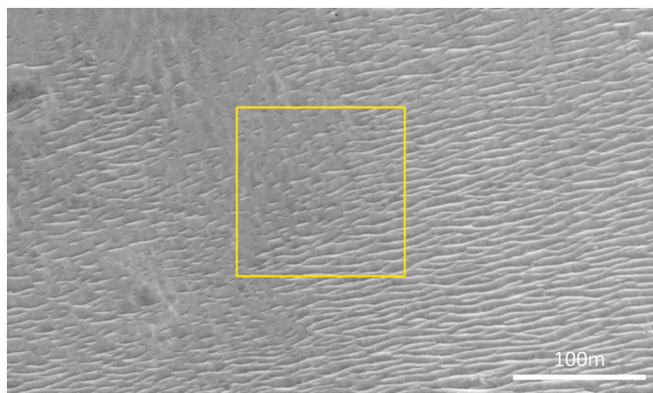


Fig. 3. Variation in the spatial density of discontinuous ripples. The yellow square is 128 m across, the size of a training framelet (HiRISE: ESP_046103_2030).

Table 2
Final ontological classes.

Surface classes	Likely traversability
Non-bedrock	
1. Smooth, Featureless	Uncertain Traversability
2. Smooth, Lineated	Low Traversability
3. Textured “Non-Bedrock”	Uncertain Traversability
Bedrock	
4. Smooth “Bedrock”	High Traversability
5. Textured “Bedrock”	Fair traversability, potential for Localised Hazards
6. Rugged “Bedrock”	Very Low Traversability
7. Fractured “Bedrock”	Localised Hazards
Dispersed Classes	
Large Ripples	
8. Simple form large ripples, Continuous	Low Traversability
9. Simple form large ripples, Isolated	Potential Low Traversability
10. Rectilinear form large ripples	Very Low Traversability
Small Ripples	
11. Continuous small ripples	Low Traversability
12. Non-continuous small ripples, Bedrock substrate	Localised Hazards
13. Non-continuous small ripples, Non-Bedrock substrate	Low Traversability
Other Cover	
14. Boulder fields	Localised Hazards

3.2.2. Class 5: textured “bedrock”

This class can confidently be interpreted as bedrock due to visible, significant small-scale structure and/or relief (even if appearing random). Textured terrains are gently scalloped, pitted, or undulating at a 5–10 m horizontal scale. This class is challenging to describe generically, since various textures can produce roughness, including pitting, craters, or the general roughness of the exposed surface. Areas with a fractured morphology form their own class, although they frequently grade into “textured bedrock” in areas where fracture patterns are weathered into hummocky ground. The traversability of textured bedrock is expected to depend on the prevalence of local hazards and relief, but should generally be fairly good.

3.2.3. Class 6: rugged “bedrock”

This class can be confidently interpreted as bedrock since it comprises areas of rugged, high-relief terrain. The crisper texture shows that it is not covered by regolith. This class can include small craters, scarp crests, and rugged hillslopes. Crater scarps are often classified as small areas of rugged bedrock, within a wider textured surface. Rugged terrain can also be found around the edges of some fracture networks, where the fracture pattern is too degraded to be recognisable as such. Rugged bedrock is expected to have poor traversability, with a large proportion

of hazards. Progress through such regions is expected to be slow.

3.2.4. Class 7: fractured “bedrock”

Fractured terrains are confidently interpreted as bedrock due to textures comprising clear linear, polygonal, or rectilinear fracture patterns. Traversability of fractured bedrock is expected to be highly situational, as fractures could provide a substantial hazard if they are deep or wide enough to block the path of the rover.

3.3. Large ripples (decametre scale)

3.3.1. Class 8: simple form large ripples, continuous

A continuous field of aeolian material where bedforms merge into one another with no evidence of bedrock or non-aeolian regolith in between. These ripples exhibit a clear ridge crest and are of a “simple” or “sinuous” morphology (Balme et al., 2008), since all the ripples are parallel. The wavelength of the features is of the scale of 5–20 m. Bedforms indicate the presence of loose, wind-deposited particles, with traversability dependent on thickness and spatial extent of the deposit. Large patches of continuous ripples would provide some of the most dangerous or time-consuming areas to traverse. The distinction between continuous and non-continuous features is thus important for determining their extent, and thus the hazard they might pose.

3.3.2. Class 9: simple form large ripples, isolated

These bedforms are isolated from one another, surrounded by bedrock or non-aeolian regolith. They have a clear ripple-like morphology with a distinctive ridge crest. Across-bedform distance >5 m. Isolated ripples would be expected to provide poor traversability, and their large size could make them challenging or time-consuming to navigate around.

3.3.3. Class 10: rectilinear form large ripples

Rectilinear ripples comprise a continuous field of >5 m scale bedforms as with type 8 above. However, the bedform crests are not parallel, but instead form a complex rectilinear or polygonal network. This class is only found as large ripples. Large rectilinear ripples could be very difficult or time-consuming to traverse since the rover could become surrounded by impassable sand ridges. Fortunately these features have a very small spatial extent, so the chance of landing in the vicinity of them is low.

3.4. Small ripples (metre scale)

All features at this scale have a simple morphology.

3.4.1. Class 11: continuous small ripples

This class comprises a continuous blanket of small ripples <5 m wide across the ridge crest. No bedrock or underlying regolith can be seen between the ripples. Continuous small ripples would present a substantial challenge to traversability, and should likely be avoided.

3.4.2. Class 12: non-continuous small ripples, bedrock

This class consists of small (<5 m across) ripple-like bedforms that dominate the surface but are not continuous. Bedrock can be observed in between the bedforms. This class is used when there are too many small bedforms to label individually, but cover is not total (>25%; <100% coverage by area). Discontinuous ripple patches are likely to be more easily traversable than continuous ones. Areas of bedrock between the ripples would provide grip for rover wheels, so long as the bedforms can be navigated around. The ripples themselves would still pose localised hazards, so progress would likely be slow.

3.4.3. Class 13: non-continuous small ripples, non-bedrock

This class is similar to type 12, but the underlying substrate is formed of non-bedrock material. The same uncertainties that apply to open

areas of non-bedrock terrain apply here, with the added complication of localised hazards from the maze of small ripples and their associated sand deposits.

3.5. Other cover

3.5.1. Class 14: boulder fields

Boulder fields consist of dense accumulations of float-rocks closer together than a few meters (on the order of ten rocks per 10 m²). Boulder fields will probably be difficult for the Rover to traverse. Some very dispersed fields could be navigable, although progress would likely be slow. Dense block fields could be impassable.

4. Data collection

Fig. 4 shows the workflow for the project, indicating which stages of the work were conducted by the computer science team (yellow) and the geomorphology team (blue). Selection of framelets is detailed in Section 4.1, while 4.2 discusses the setup of the NOAA-H Dataset Annotation Tool (DAT). Labelling is detailed in 4.3 and the training and evaluation of the model in Sections 5.1–2 and 5.3–6 respectively. Section 6 then details the geomorphological assessment of the output rasters.

4.1. Framelet selection

100 HiRISE images with full 25 cm/pixel resolution were identified. At each landing site, the 50 images which provided the best coverage of the 1-sigma landing ellipse were selected, starting at the centre and working out. Some images were disqualified, either because they had 50 cm/pixel resolution, or because they formed half of a stereo pair with an image which had already been selected, and so covered exactly the same features. The choice of which image from each pair to include was random.

Once the best possible coverage of the 1-sigma ellipse was attained, outlying images, scattered across the 3-sigma ellipse were added, until there were 50 for each site. These were chosen to provide additional examples of classes which were underrepresented in the preceding images.

Selection of training data took place in summer of 2018, so the landing ellipses used to define the study areas date from that time. The position of the Oxia Planum ellipse has subsequently changed due to the revised ExoMars mission timeline, while Mawrth Vallis was not ultimately selected. Many more HiRISE images have been acquired for these sites in the time since this project began. However, in 2018 the 1-sigma landing ellipses at both sites already had almost full coverage of red-band HiRISE images.

A series of 128 m × 128 m framelets were digitised using ArcGIS at locations where representative features were found. These were then exported to a non-georeferenced GIS shape file, which was used to automatically crop out the relevant sections of the HiRISE images, and upload them to the NOAA Data Annotation Tool (DAT) (Section 4.2).

Between ten and twenty framelets were selected for each HiRISE image, giving a total of 1504, covering both landing sites as shown in Table 3 and Fig. 5. An equal number of images were surveyed from each study area, although slightly more framelets were ultimately selected from Mawrth Vallis than Oxia Planum.

Framelets were placed manually so as to ensure that their contents provided the maximum number of examples. It was vital that the training dataset provided a representative catalogue of all ontological classes, and the full variety of intra-class variations present at the sites. Random selection and placement would not have yielded sufficient variety. Terrain with large spatial extents would have been disproportionately represented, while rare features would not have been sampled at all. Type examples were selected which demonstrated both “good” examples of the various classes, and less distinct morphologies which still conformed to a class. The aim was to ensure that every terrain the

algorithm encountered could be assigned a class.

Very few framelets were directly adjacent. This was only done in the case of very rare morphologies such as rectilinear ripples, where it was essential to get as many framelets from a given example as possible. In most other cases they were located hundreds of metres apart. Using discontinuous framelets allowed examples to be drawn from a much wider area than if a single HiRISE image had been labelled in its entirety. This ensured that the algorithm had a representative suite of examples to learn from and makes the resulting model more transferable between different parts of the study area. The use of framelets also made the labelling task more manageable since the labeller is shown small sections of image in turn, rather than having to segment a single large area. It also allowed the DAT to be built on top of the existing “Zooniverse” platform. This division of data was only used as part of the training and validation process. The final model classified entire HiRISE images and not just small subsections.

Labelling was conducted in two batches. An initial set of training data was labelled early on in the project. This was later supplemented with additional images, both to increase overall support, and to allow the balance between the different classes to be improved by targeting features which were under represented the first time around. This allowed the computer science team to attempt a first run of the model with the initial batch, which provided useful information as to which classes required more support. Different framelets were reserved for validation in each stage of the experiment, to avoid overfitting. The images used, and the number of framelets contributed by each image are listed in Table 3.

The training dataset is representative of the variety of textures present at the sites, but not their relative proportions. Some landform types are much more common in the study areas than others. Examples of the “textured non-bedrock” class occur in almost every framelet acquired, and “rugged bedrock” is also extremely common. “Ripples” are common in most areas, but vary considerably in scale and distribution. Classes such as “lineated non-bedrock”, “smooth bedrock”, “boulder fields”, and “rectilinear ripples” are much rarer. Uncommon landforms were deliberately targeted, and so occupy a larger proportion of the framelets from the images in which they do occur. Even so, there are fewer examples of them, so while every effort was made to maximise coverage of these features, they inevitably provided a smaller area of labelled pixels overall. The number of pixels labelled as each ontology is shown in Fig. 8. There is a large difference in support between the more common landforms, and those which are rare. Abundance of a class in the training dataset is not representative of its abundance across the site as a whole, since the surface area of common classes is so much higher than that of rare ones that the dataset would not be usable if the actual proportions were used.

4.2. The dataset annotation tool (DAT)

Features within the framelets were labelled using the NOAA Dataset Annotation Tool (DAT) (Wallace et al., 2017), producing a dataset of pixel-class pairs. In contrast to the initial NOAA project, labelling was carried out by the research team, rather than through a citizen science program. The degree of expertise required to distinguish between subtly different classes using remote sensing data precluded a crowd sourcing approach in this case.

The DAT originally proposed in Read et al. (2018) was extended to enable manual annotation and pixel level labelling of HiRISE images based on the defined ontological classes. It was designed on top of the Oxford University Zooniverse platform (Simpson and De Roure, 2014). Zooniverse provided extensive built-in functionality, but certain features needed to be extended in order to ensure high quality annotation data, and to support the users during the labelling campaign.

Newly added features included a contextual ‘zoomed-out’ version of the image, and showing metadata in the participant’s view. Image number of the original HiRISE data product was provided, and the

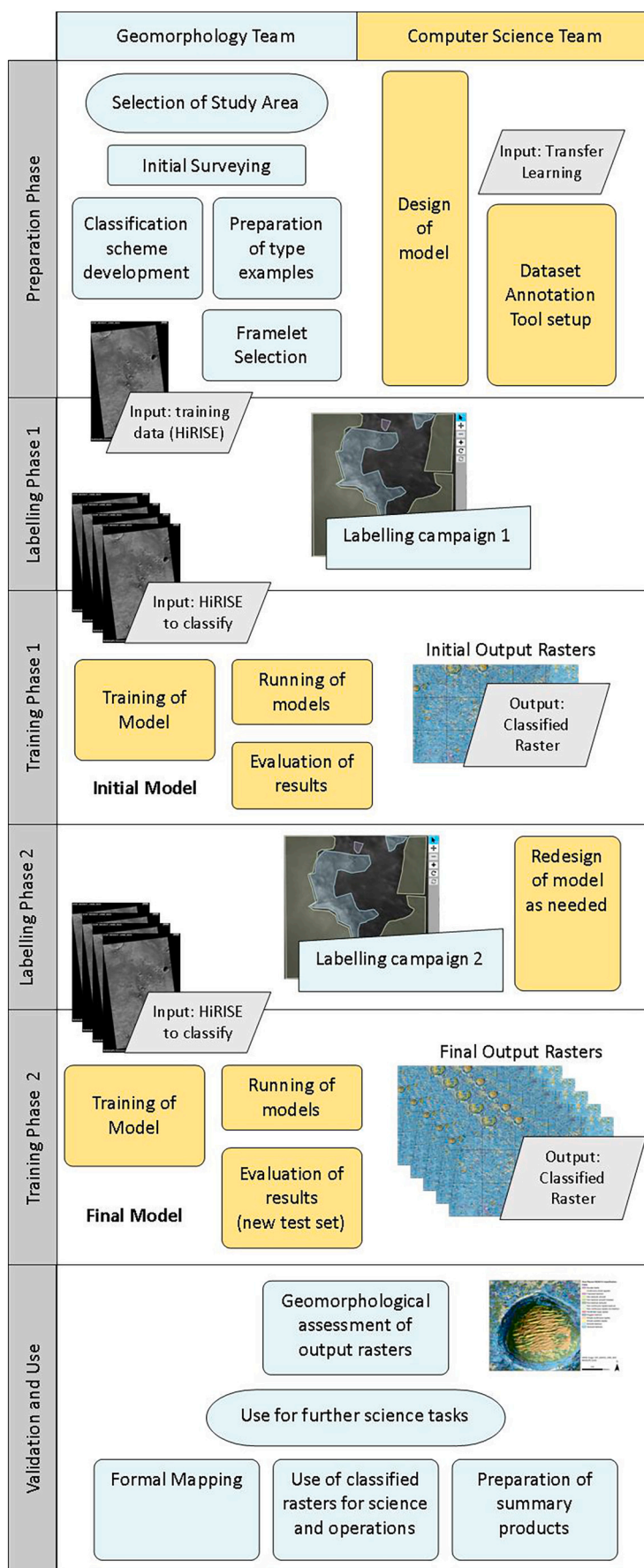


Fig. 4. Illustration of project workflow, and the roles played by the Computer Science and Geomorphology teams.

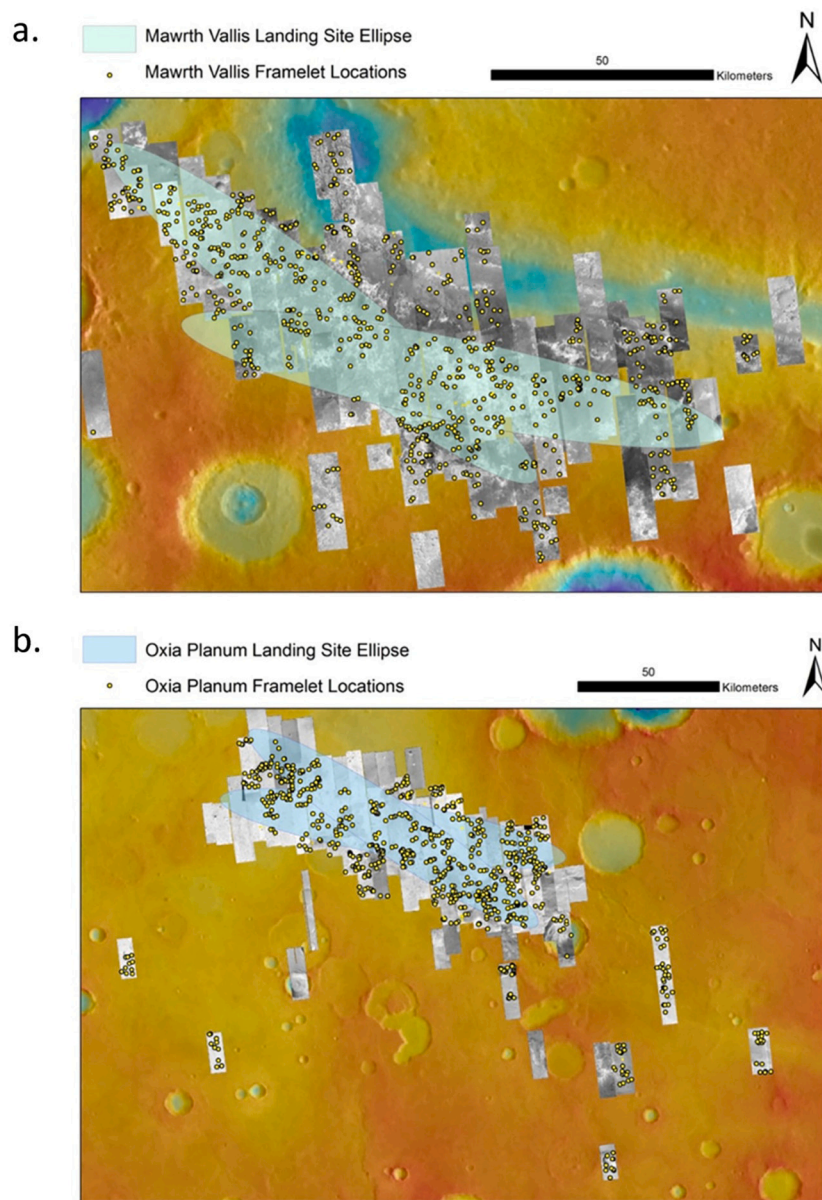


Fig. 5. Maps of a. Mawrth Vallis and b. Oxia Planum, showing distribution of framelets and HiRISE images (McEwen et al., 2010, NASA/JPL/UoA) over the MOLA global topographic map (Smith et al., 2001).

locations of the top left and bottom right corners of the image were shown in both pixel coordinates and latitude and longitude. This allowed further analysis of the image to be conducted using GIS software in parallel with labelling in the DAT. Fig. 6 shows an example of the interface, including the contextual zoom and metadata display. A global progress indicator was provided allowing users to visualise overall annotation progress by seeing pins on a map of Mars.

4.3. Labelling procedure

Each framelet was briefly examined to determine which classes were present. Labelling was conducted by segmenting the image using a polygon tool, ensuring that each polygon only included terrain of a single class. This polygon was then assigned the appropriate label, producing a vector dataset indicating which pixels shared a class label. The aim was always to ensure that all labelled pixels conformed to the chosen class.

In some cases, an area was segmented in several blocks, so as to more

effectively define its extent. More complex, or harder to interpret regions were segmented last, carefully building shapes around areas which had already been defined. In cases where several small blocks were contained within the extent of a larger area, these were segmented first. This allowed the user to more effectively draw a larger shape around them. Transitional zones were frequently left blank, or classified later, when experience drawing the earlier segments resulted in an improved understanding of the image. In cases where longer ripples overlaid bedrock and non-bedrock material, the boundaries of the surface blocks formed the boundaries of the labels. Fig. 7 illustrates the process of digitising framelets in the DAT. Before and after images are presented, showing how the different terrains are identified.

Polygons were drawn as tight around the features being digitised as possible. In instances where boundaries were not clearly defined only the central, definitive block was segmented. In rare cases where an overlap occurred, the polygon which was drawn last took precedence. Every effort was taken to avoid overlaps. They occasionally occurred due to user error but not often enough to be statistically significant.

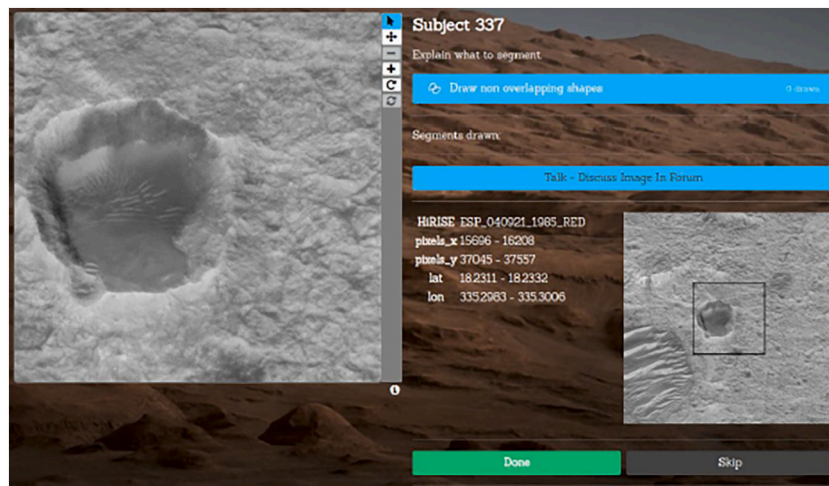


Fig. 6. The Dataset Annotation Tool. (HiRISE image: ESP_040921_1985).

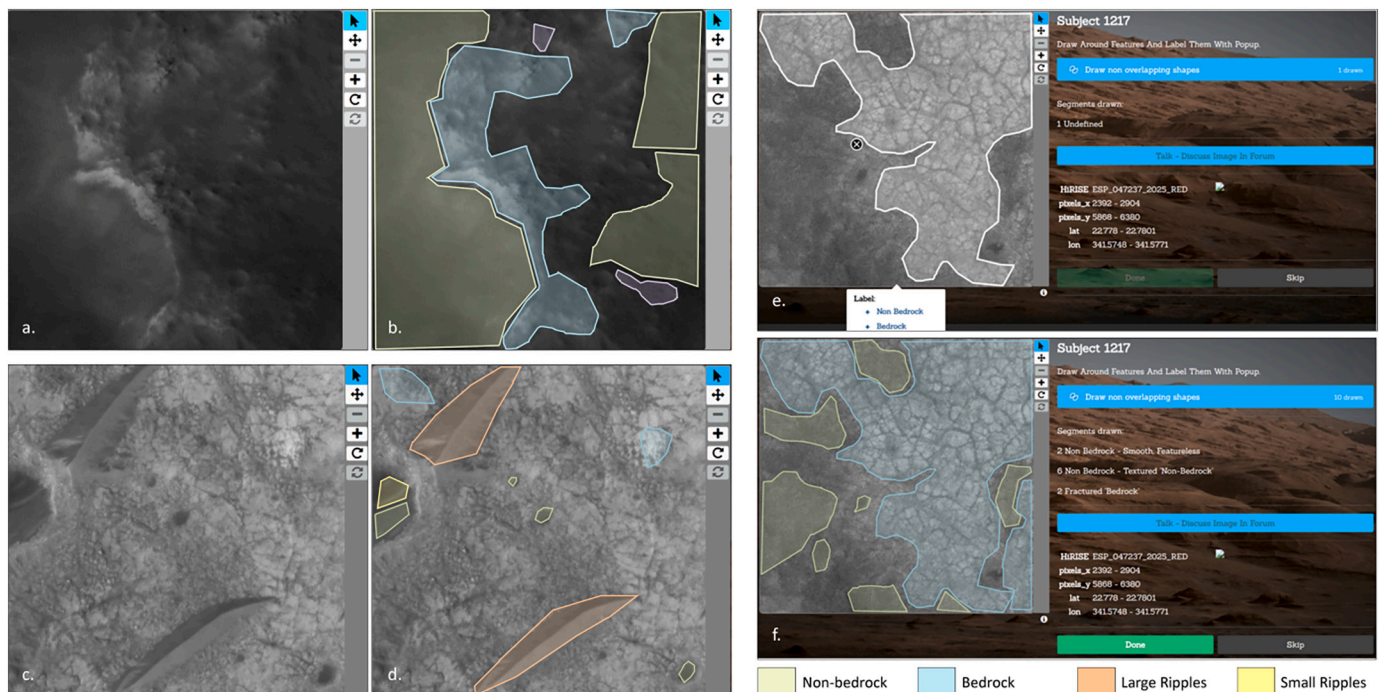


Fig. 7. a–b. Digitisation of a block of rugged terrain, showing the “buffer zones” of unclassified ground at the transitions between the classes. c–d. Labelling of large isolated ripples over fractured bedrock. The ripples and blocks of featureless ground are digitised first; the remaining area can then be labelled as fractured bedrock. The majority of the unlabelled area of the image was subsequently classified as bedrock. e–f. Process of digitising a varied area (HiRISE: ESP_047237_2025).

Not all pixels in a given framelet were labelled. It was not always possible to assign a unique class to every section of an image. The process of neatly drawing polygons around specific features sometimes precluded the possibility of making them tessellate exactly. It was often pragmatic to leave small areas blank if that made the digitisation of larger areas easier. Ensuring that the labelled areas were definitive was more important than labelling every possible example in a framelet.

Every effort was taken to ensure that the labelled areas did not just comprise “easy” or “good” examples. Labelling was conducted by a trained geomorphologist with years of experience of interpreting the martian landscape at HiRISE scale. Extensive work was carried out to determine which terrains would be included in each of the classes, and type examples were chosen which represented every expression of a given class, ranging from very good examples to extremely poor ones. Thus the only sections which were not labelled were ones where the

terrain was truly equivocal, and it was not possible to make a reliable or repeatable determination. This included some transitional regions where the terrain conformed to the characteristics of multiple classes.

It was ultimately decided that a single geomorphologist should conduct all of the labelling. This reduced the effect of subjectivity, and ensured that the images were labelled consistently. We acknowledge that this means that our algorithm will only be as good as the geomorphologist who trained it. In principle a larger number of labellers could have provided an additional check, and given the algorithm a more varied perspective. However there is also a chance that additional labellers would have introduced additional error and subjectivity into the method. It would be interesting to compare the two approaches, but doing so was beyond the scope of the present work. The constraints of the project and the team’s time meant that it was not possible to employ multiple labellers in the time allotted for the project.

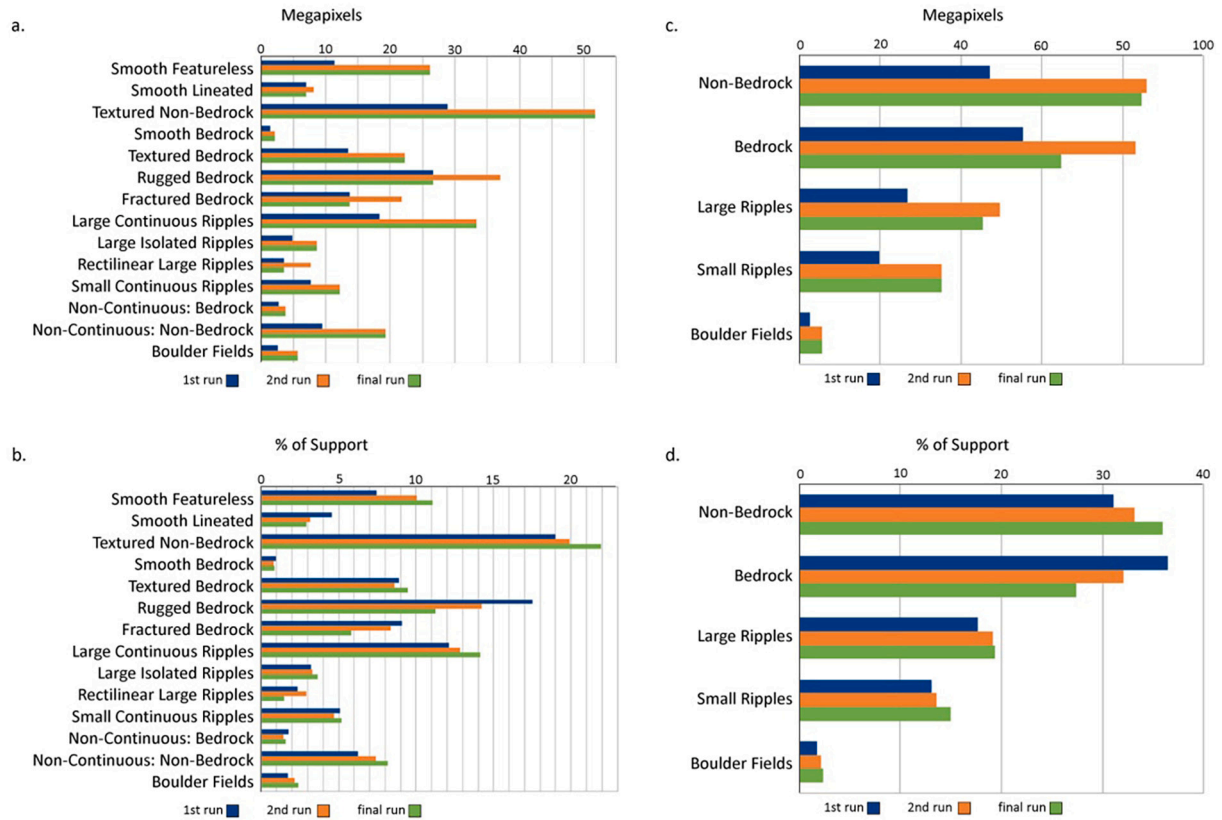


Fig. 8. Distribution and support of labelled pixels for; a&b. All classes and c&d. interpretive groups.

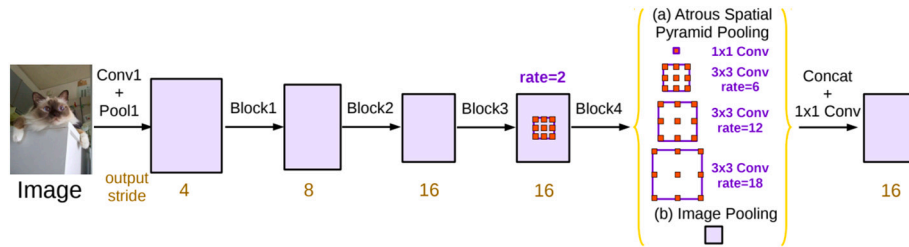


Fig. 9. Illustration of Atrous convolution, in the Google DeepLab model. Output stride is the ratio of the original image and the output feature map. (Image credit: Chen, 2017).

A total of $\sim 2.36 \times 10^8$ pixels were labelled across the 1504 frame-lets. The labelling work was split into two campaigns, to allow the DNN performance to be evaluated before deciding what additional labelling was required.

5. Training the DNN

5.1. Dataset

Two labelling activities were conducted as part of the NOAH-H Project. The “First Run” included only the initial batch of labelling. The “Second Run” included data from both batches, ensuring an overall increase in support (i.e. number of pixels classified) for the various classes (Table 4). A third dataset known as the “Final Run” balanced the overall support for some classes by making small but focused modifications to the second run. In order to properly evaluate the system, the dataset was randomly split into two sets: 90% for Training and 10% for Validation. A unique validation set was created for each campaign to avoid overfitting. In addition to these unique validation sets, a separate

“test set” was produced, which was used to compare the results of the three runs.

To evaluate the performance of the NOAH-H system as a broader semantic segmentation network, the ontological classes were also combined into the five interpretive groups; Bedrock, Non-Bedrock, Large Ripples, Small Ripples, and Boulder Patches. Fig. 8 presents a full breakdown of the number of labelled pixels and the support for each class in the total number of pixels, i.e. the percentage of data of each class in the whole dataset (including both training and validation sets).

The number of MP across all classes almost doubled between the first and second runs. The percentage of support increased in the classes with low numbers of labelled pixels and became more even for the two dominant classes “Non-Bedrock” and “Bedrock”.

5.2. DNN methodology

The aim of the NOAH-H project was to address the identification of semantically accurate predictions and the definition of segmentation maps along object boundaries of novelties/anomalies from HiRISE

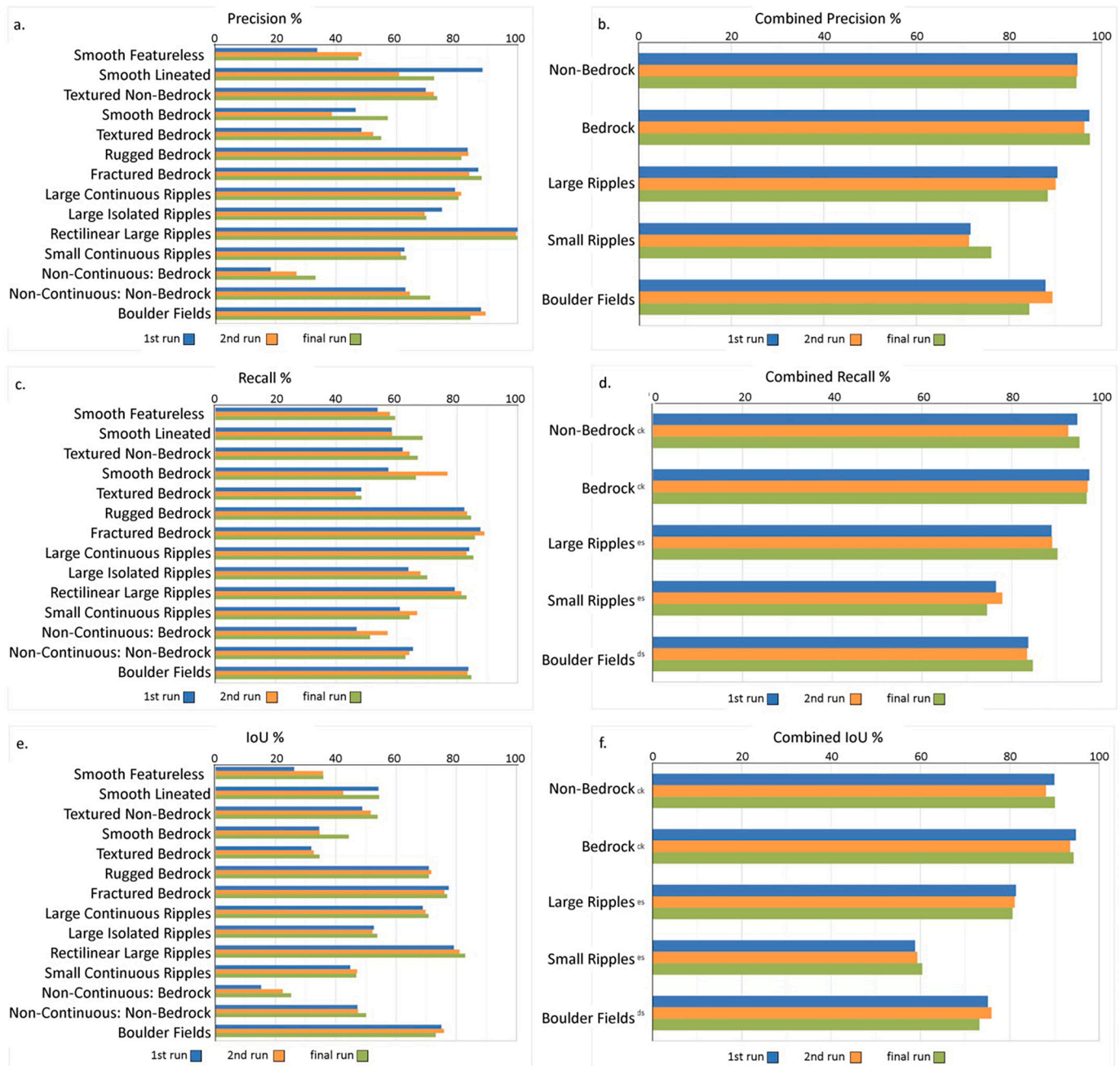


Fig. 10. Comparison of precision (top), recall (middle), and IoU (bottom) across all runs. Descriptive classes (left), and interpretive groups (right).

images.

To overcome this challenge, we focused on a technique known as semantic segmentation to process small areas of the HiRISE image by extracting local features. The principle of semantic segmentation models (Badrinarayanan et al., 2017; Chen et al., 2018b; Papandreou et al., 2015) involves classifying each pixel of an image by analysing the region around it, often called the receptive field, and passing it through a deep network to compute a class prediction. A common state-of-the-art semantic segmentation approach has been described in Long et al. (2015), in which 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 (Chen, 2017; Chen et al., 2018a; Liu et al., 2019) the architecture of

which is shown in Fig. 9.

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. It offered an efficient mechanism to control the field-of-view and found the best trade-off between accurate localisation (small field-of-view) and wider range context with more semantic information (large field-of-view).

Atrous convolutions are particularly suited to dense semantic segmentation as they expand the receptive field without losing resolution or coverage. There are cases where one wants to balance the pixel level accuracy such as detection around the edges of a feature and information of a wider context. To solve this problem, multi-scale convolutional layers are used at a cost of efficiency. To increase the runtime

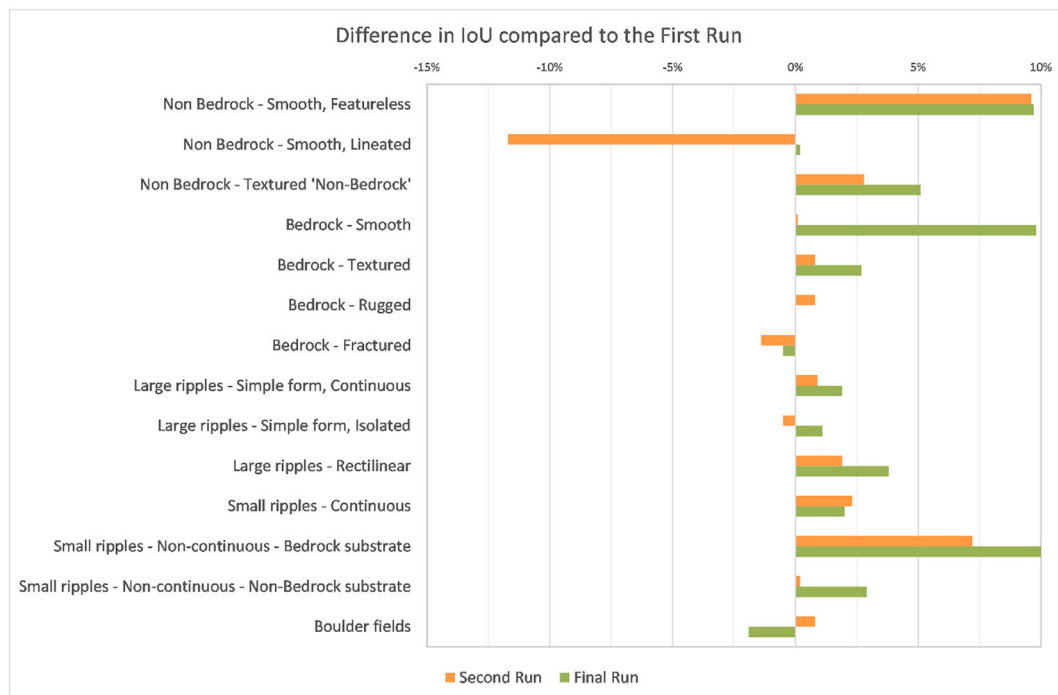


Fig. 11. Difference in IoU of the “Second” and the “Final” run.

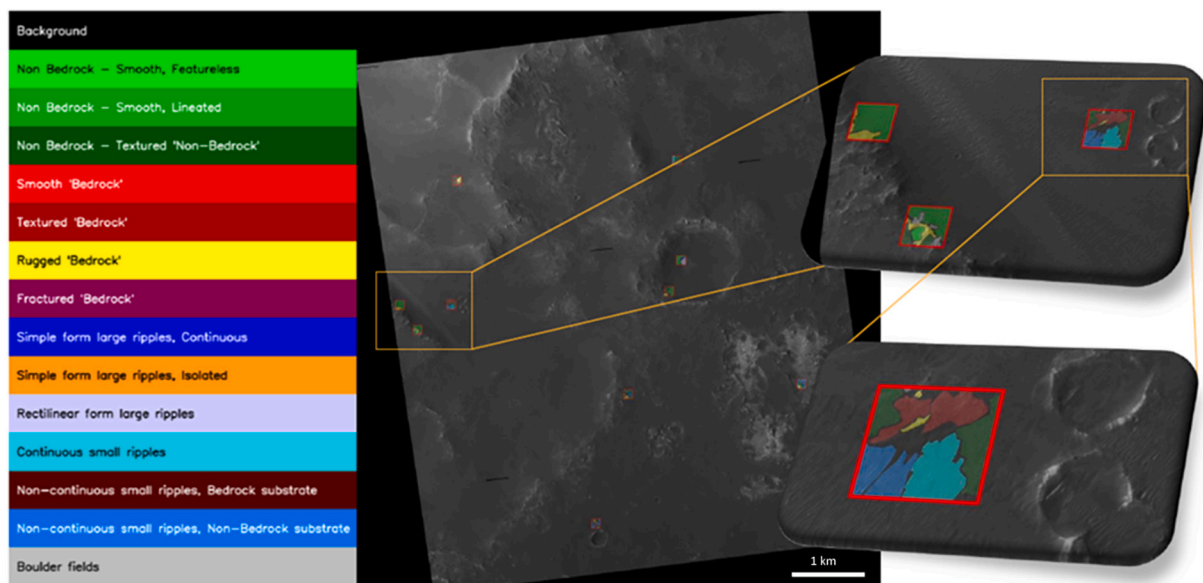


Fig. 12. Example of a labelled HiRISE image with colour codes for all classes. This illustrates how labelled framelets appear, and how they are distributed on a representative image. (HiRISE image: ESP_019809_2030_RED).

performance atrous convolutional layers are used instead of the multi-scale ones as shown in Yu and Koltun (2016). They were thus a good choice for this study.

The training of the model was done on a Linux machine using Python and Tensorflow for ease of adaptation and visualisation. Input size was the entire framelet padded with an extra pixel at the edges if required. The initial run of the model was allowed to train for 30,000 iterations with a batch size of 10. Since we did not know how the model was going to behave during training this was preferable to having it automatically stop. After observing the performance, in order not to overfit, we chose to use an early-finished model that was trained to 19,500 iterations. We decided to stop any further training when we did not observe any

increase in accuracy during training, and a drop of accuracy in the remaining 10% of the training data.

5.3. Evaluation measures

The accuracy of the model was assessed in terms of the agreement between the model prediction and the manually labelled validation set. This was evaluated using the Precision, Recall and Intersection over Union (IoU) metrics. These are some of the most common metrics for semantic segmentation and are extremely effective at deciding whether a prediction is correct with respect to an object or not. They compare the model output to the original labelled framelets reserved for validation.

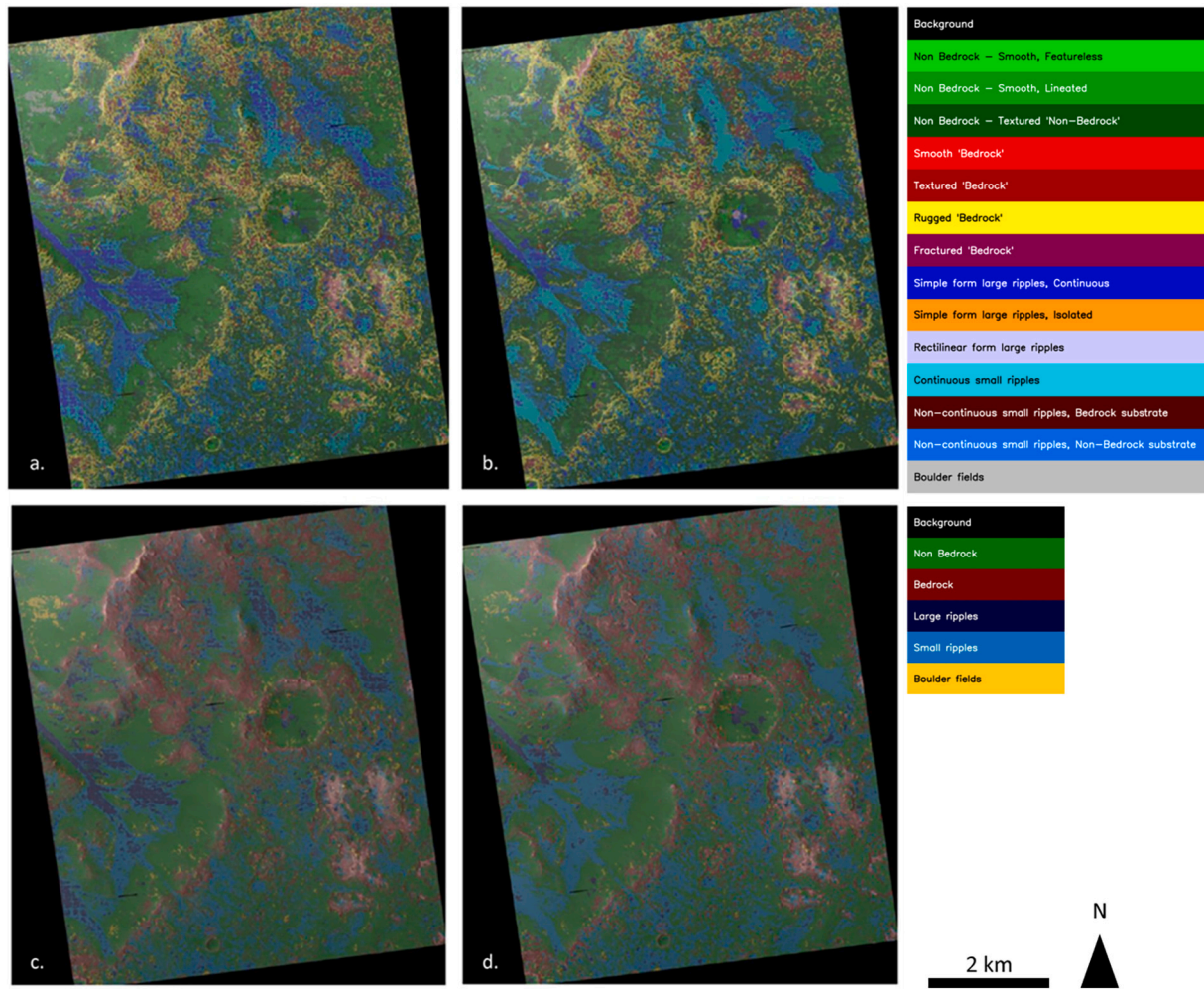


Fig. 13. Example output of the NOAH-H system, showing how the full image is classified. a. Output of the “First Run” model with all the classes of the ontology. b. Output of the “Final Run” model with all classes. c. Output of the “First Run” model with all the combined classes. d. Output of the “Final Run” model with combined classes. (HiRISE: ESP_019809_2030).

The level of support for each class is factored into the calculation of IoU. It thus provides a robust means of assessing the reliability of the model. A confusion matrix (e.g. [Tharwat, 2018](#)) was used to calculate the metrics for the individual classes.

The Precision, Recall, and IoU for each class was measured using the following formulae:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{IoU} = \text{TP} / (\text{TP} + \text{FP} + \text{FN}) \quad (3)$$

Where:

- True Positive (TP): Number of pixels correctly classified.
- False Positive (FP): Number of pixels incorrectly classified.
- False Negative (FN): Number of pixels incorrectly not classified.

Precision calculates the percentage of pixels that the AI identified with the correct label, out of all the pixels that the AI labelled as that specific class. This shows the extent to which the NOAH-H classification matched how the terrain was labelled by the expert. Intuitively, this means how well the AI did when predicting a pixel label.

Recall assesses whether the machine learning algorithm is reliably identifying all areas which were manually labelled as a specific ontology

by the science team. It calculates the percentage of labelled pixels from the evaluation dataset which were labelled correctly by the model. Intuitively, this means how well the AI did in not missing the pixels that the experts labelled.

These metrics allow the results to be considered in two key ways:

1. When the model has found a class, has it got it right? (precision)
2. When a class is known to be present, has the model found it? (recall)

The IoU essentially combines the precision and recall metrics into one. It is used to compare multiple runs or models, since only one number per class needs to be considered. Precision, Recall and IoU results are shown in [Fig. 10](#). In order to further compare various models, without looking into the individual classes, the mean IoU was calculated as an average across all IoUs for all classes. Mean IoU was weighted by the prevalence of those classes to provide a fair comparison. This gave us a single value of performance for each model, which we could easily compare.

5.4. Variation between runs

For the First Run the mean IoU across all pixels for the full class list was 72.73% while for the combined groups it was 92.5%. Because the performance of the model was higher when combining the classes, we conclude that more training data are needed for the confused classes. As

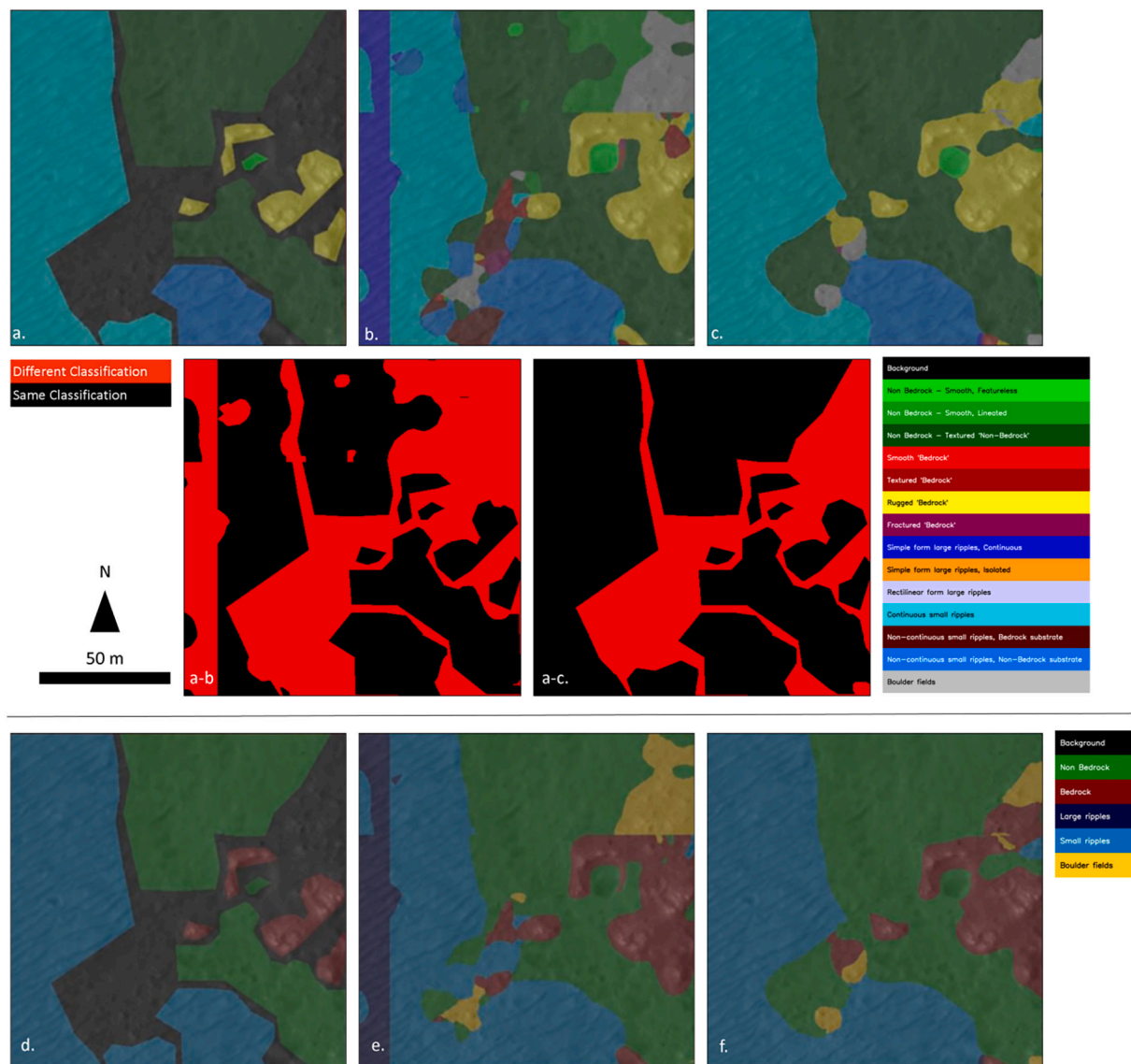


Fig. 14. Comparison example of the NOAH-H system for a 512×512 pixel (128×128 m) framelet from the validation set. a. Expert Labelled framelet showing all classes. b. Output of the “First Run” model. c. Output of the “Final Run” model. a-b. Difference between expert labelled framelet and first run. a-c. Difference between expert labelled framelet and final run. The bottom row shows the effect of combining classes into thematic groups. d. Expert labelled. e. first run. f. final run. (Hirise image: ESP_019809_2030).

shown in Table 5, more data were labelled for the Second Run, for which the mean IoU across all pixels was 72.83% (+0.09%) for the full class list and 91.91% (−0.59%) for the groups.

Despite the big increase in labelled data, the agreement with the test set remained almost the same. We increased some of the classes that were originally confused; however this caused other classes to drop in accuracy, yielding no overall gain in performance. For the “Final Run” we decided to include only the new data for the classes with low accuracy, the mean IoU across all pixels became 74.15% (+1.33%) for the full class list and 92.33% (−0.17%) for the groups.

We can see an improvement in the classes to which more training data were added, but this still caused a drop in accuracy for other classes. Fig. 11 plots the difference in accuracy of the latter two runs against the first, to better visualise the changes. This demonstrates that a better understanding of the training data is crucial to the accuracy of the model, since 11 out of the 14 classes improved in the Final Run and only 3 out of 14 decreased in accuracy compared to the Second Run.

For the groups the results were almost identical. By carefully adding the correct data to improve the training set distribution, we were able to

improve specific classes such as the small ripples when a new model was trained with the revised data. However, careful investigation is needed to not cause any decrease in other classes’ accuracy. These results show such a decrease in the combined classes and the individual classes for smooth and lineated non bedrock, where there was at least an 11% decrease in performance.

The modification of the training data between the “Second Run” and “Final Run” was informed by the results of the previous attempt and showed that data balancing techniques needed to be included in future models. The total number of images should not have been selected from the start of the project but rather by checking the balance of the classes and the confusion between specific classes. By improving the balance of the training dataset for future versions of the model, we were able to improve the results, without overfitting to the specific sites.

5.5. Variation between classes

When considering whether the results are fit for purpose, it is important to analyse the substantial variation in precision, recall, and

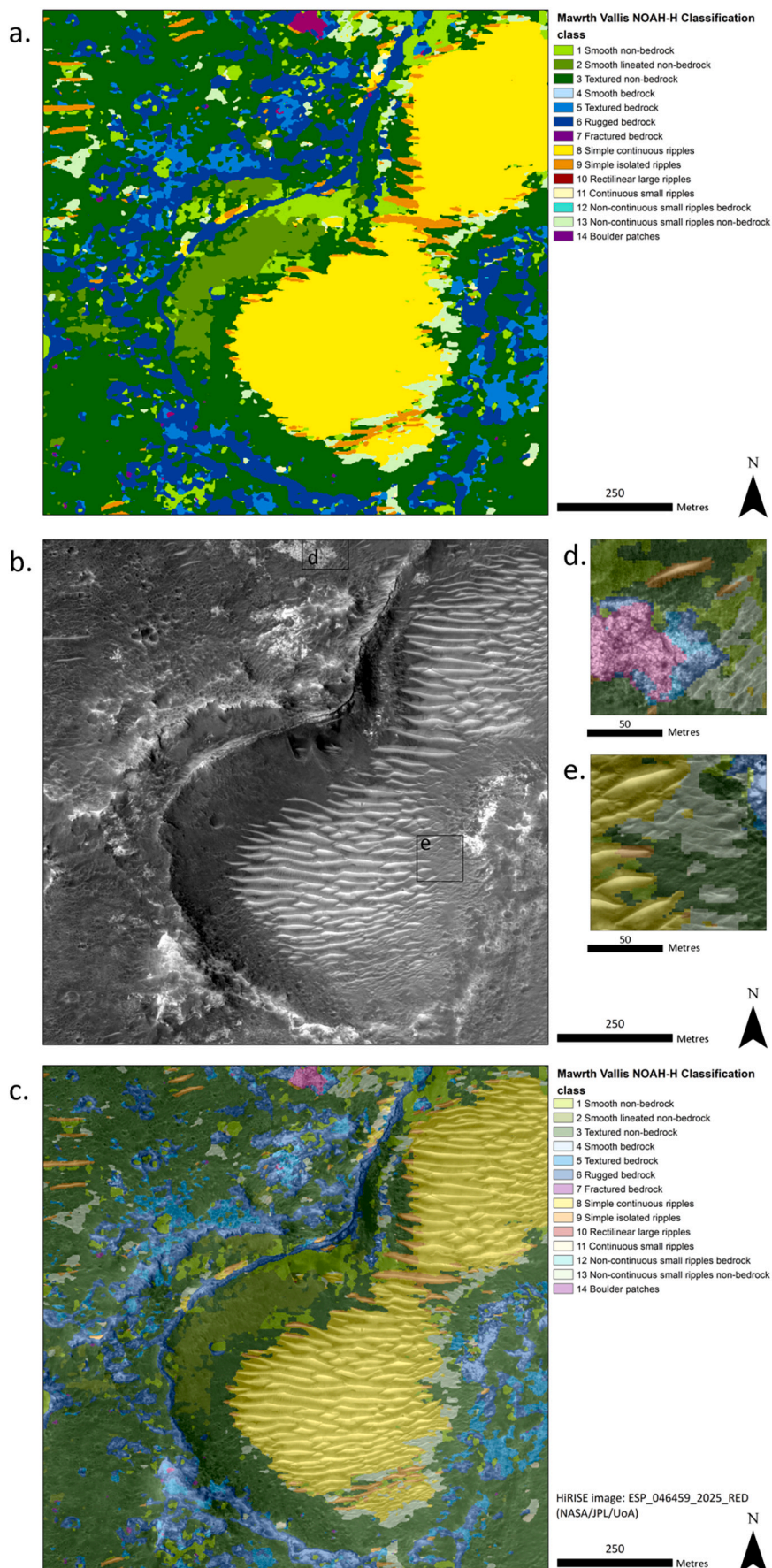


Fig. 15. Double crater in Mawrth Vallis with large field of aeolian ripples. Crater rim correctly classified as rugged bedrock (dark-blue), much of the ejecta is also rugged, with textured (mid blue) areas. A small patch of fractured bedrock (magenta) is correctly identified to the north of the image. The dense region of large continuous ripples (bright-yellow), with large isolated ripples (orange) at the periphery is detected, as are other isolated ripples outside the craters. All have been delineated well. The contact between the regions of large and small bedforms has been correctly identified. A large region of smooth lineated non-bedrock (pale-green) is correctly identified on the inner crater wall. The edges of the ripples are very clearly identified, and the transitions between the different bedrock and non-bedrock classes are largely accurate. a. NOAH-H output raster, b. Original HiRISE image from ESP_046459_2025. c. NOAH-H over HiRISE. d. close up of fractured ground. e. close up of aeolian bedforms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

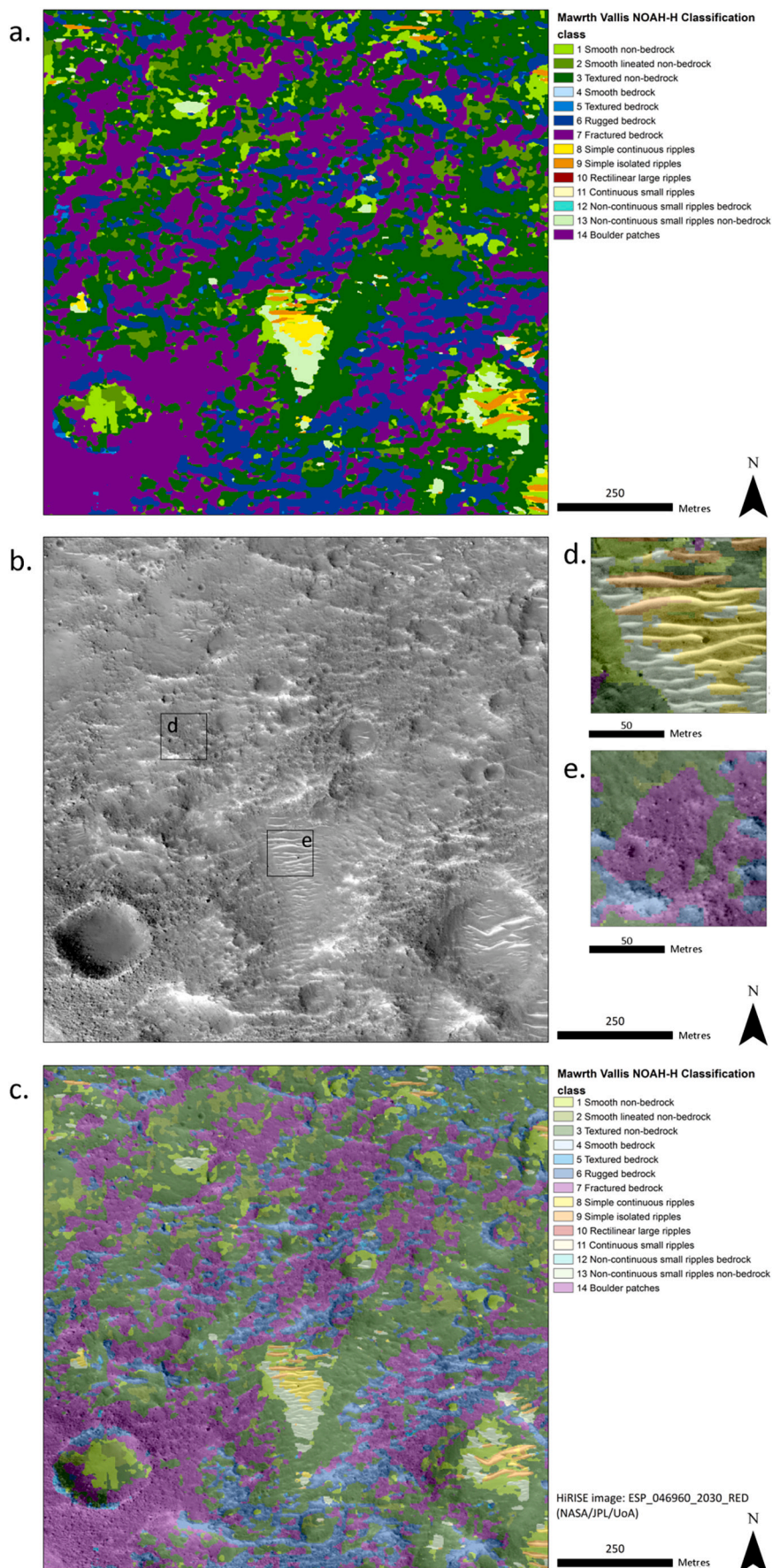


Fig. 16. Boulder field (purple) between several small impact craters interspersed with small ridges of rugged bedrock (dark blue). Ground not covered by boulders consists of a textured non-bedrock surface (green). The classified raster highlights these features very well. In most cases the boundary between discontinuous small ripples over non-bedrock (turquoise) and continuous ripples (pale yellow) has been identified appropriately. However, it does cut across some individual bedforms. a. NOAH-H output raster, b. Original HiRISE image from ESP_046960_2030. c. NOAH-H over HiRISE. d. close up of aeolian bedforms. e. close up of boulder field. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

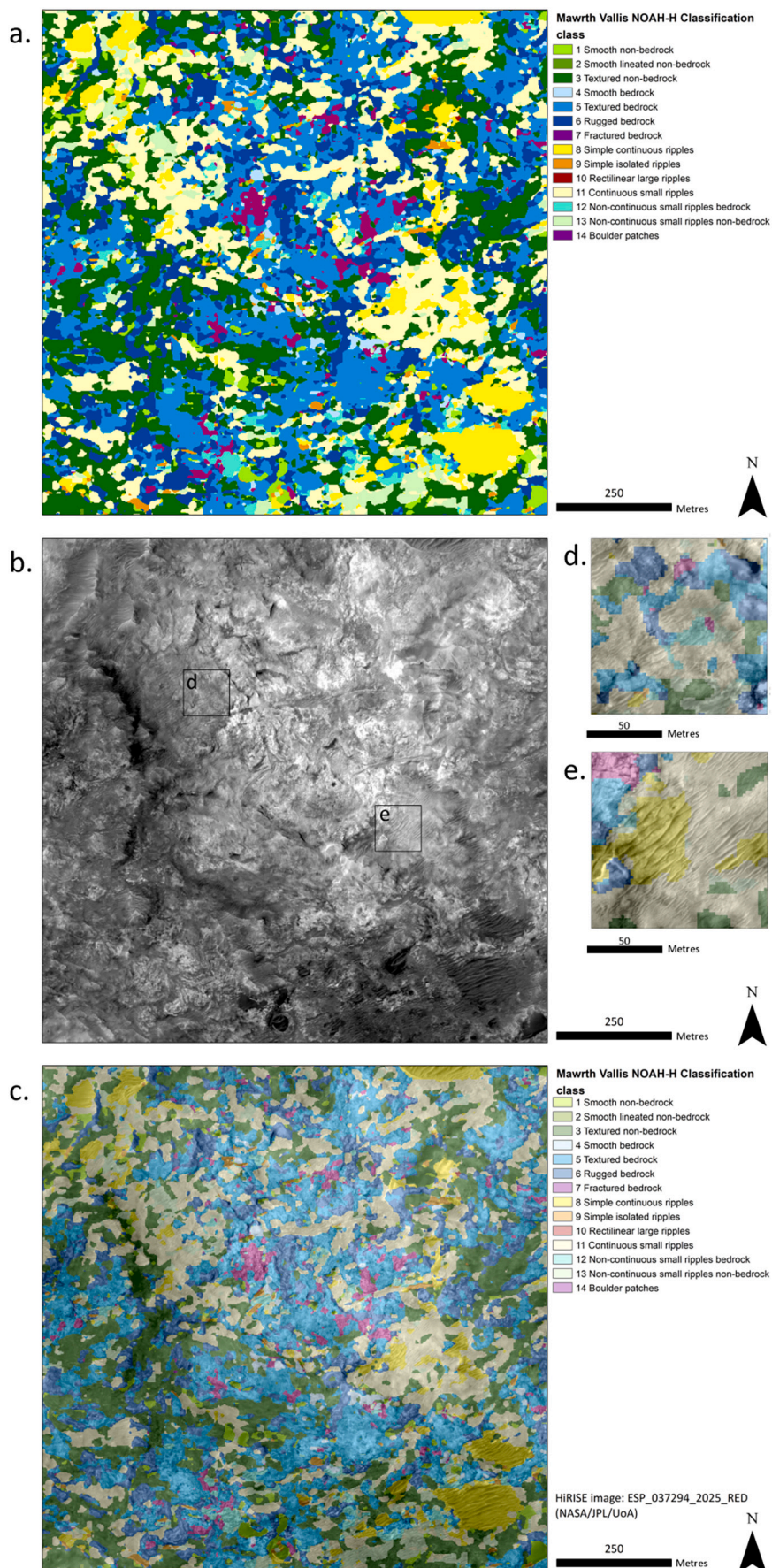


Fig. 17. Rough terrain in Mawrth Vallis, with small patches of ripples. A few large isolated ripples (orange) are scattered across the site, either as part of larger regions of small ripples, or individually. Small patches of fractures (magenta) are found reliably within large areas of textured and rugged bedrock (mid & dark blue). The edges of ripple classes are well defined, and the transitions between different types of ripple are largely correct. Small areas of interspersed bedrock and non-bedrock are generally identified correctly; however, the transitions between areas of different textures are not always precise, since the margins are actually gradational. **a.** NOAH-H output raster, **b.** Original HiRISE image from ESP_037294_2025. **c.** NOAH-H over HiRISE. **d.** close up of varied area. **e.** close up of aeolian bedforms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

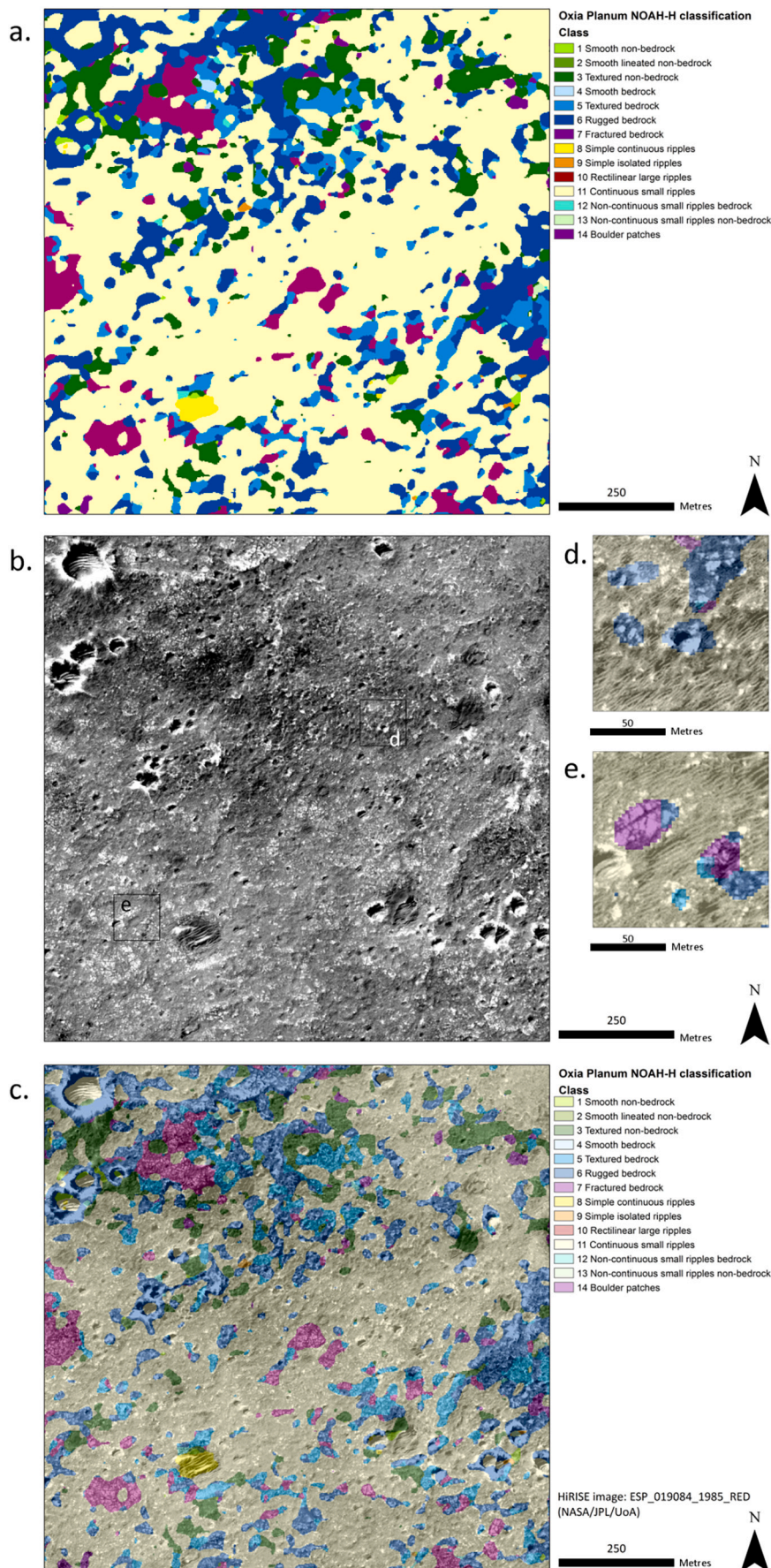


Fig. 18. Blanket of meter-scale continuous or near-continuous ripples. These are hard to identify unless the image is viewed at full-resolution and, as they form a relatively thin mantle, they do not obscure the underlying relief when viewed at lower image resolution. The ripples overlie a variety of rugged, textured, and fractured bedrock areas. The borders between the ripple-covered areas and the exposed bedrock are clearly delineated, although there are a few cases where small ripples extend into areas classified as textured non-bedrock. The NOAH-H classification very clearly shows the large scale distribution, even if the accuracy of specific margins can be debated. **a.** NOAH-H output raster, **b.** Original HiRISE image from ESP_019084_1985. **c.** NOAH-H over HiRISE. **d.** close up of small ripples in varied area. **e.** close up of fractured ground.

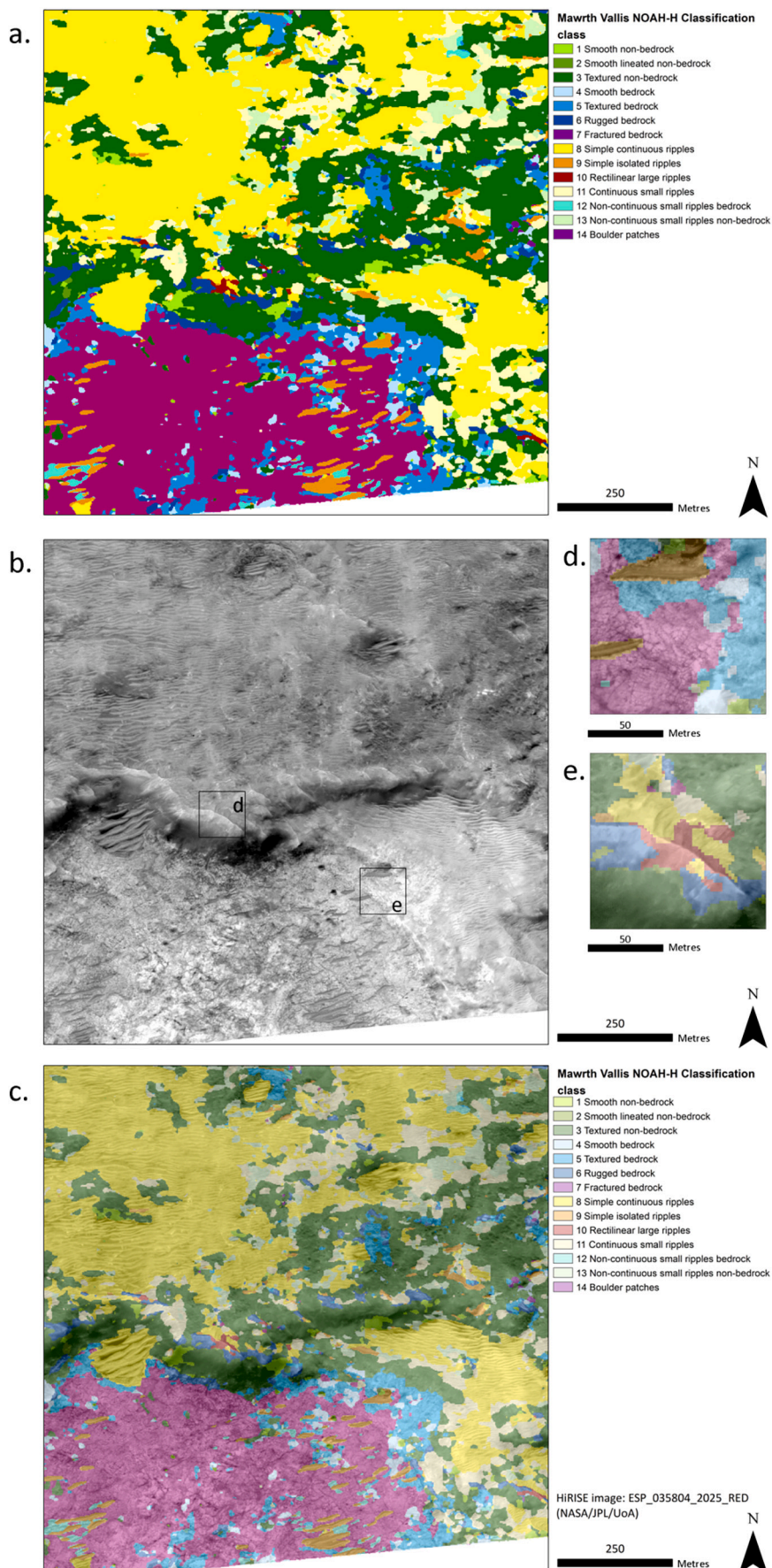


Fig. 19. Large field of ripples and areas of fractured bedrock in Mawrth Vallis. Variations in ripple morphology in the northern half of the image are well defined in the NOAH-H output, with transitions between large (mainly continuous) and small (usually discontinuous) ripples being clearly identified. Large isolated ripples (orange) are very clearly segmented. The large expanse of fractured ground in the south (Magenta) contains various small patches classified as other terrain types, mainly from three bedrock classes. In most cases these do correspond to non-fractured areas on the margin of the fractured domain. However, there are a few small areas (generally a few tens of pixels across) where subdued fractures are still present, but not classified as such. The few areas where discontinuous ripples (turquoise) overlie the fractured bedrock are generally identified correctly. Many of these border the larger, isolated ripples. a. NOAH-H output raster, b. Original HiRISE image from ESP_035804_2025. c. NOAH-H over HiRISE. d. close up of fracture patterns. e. close up of varied area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

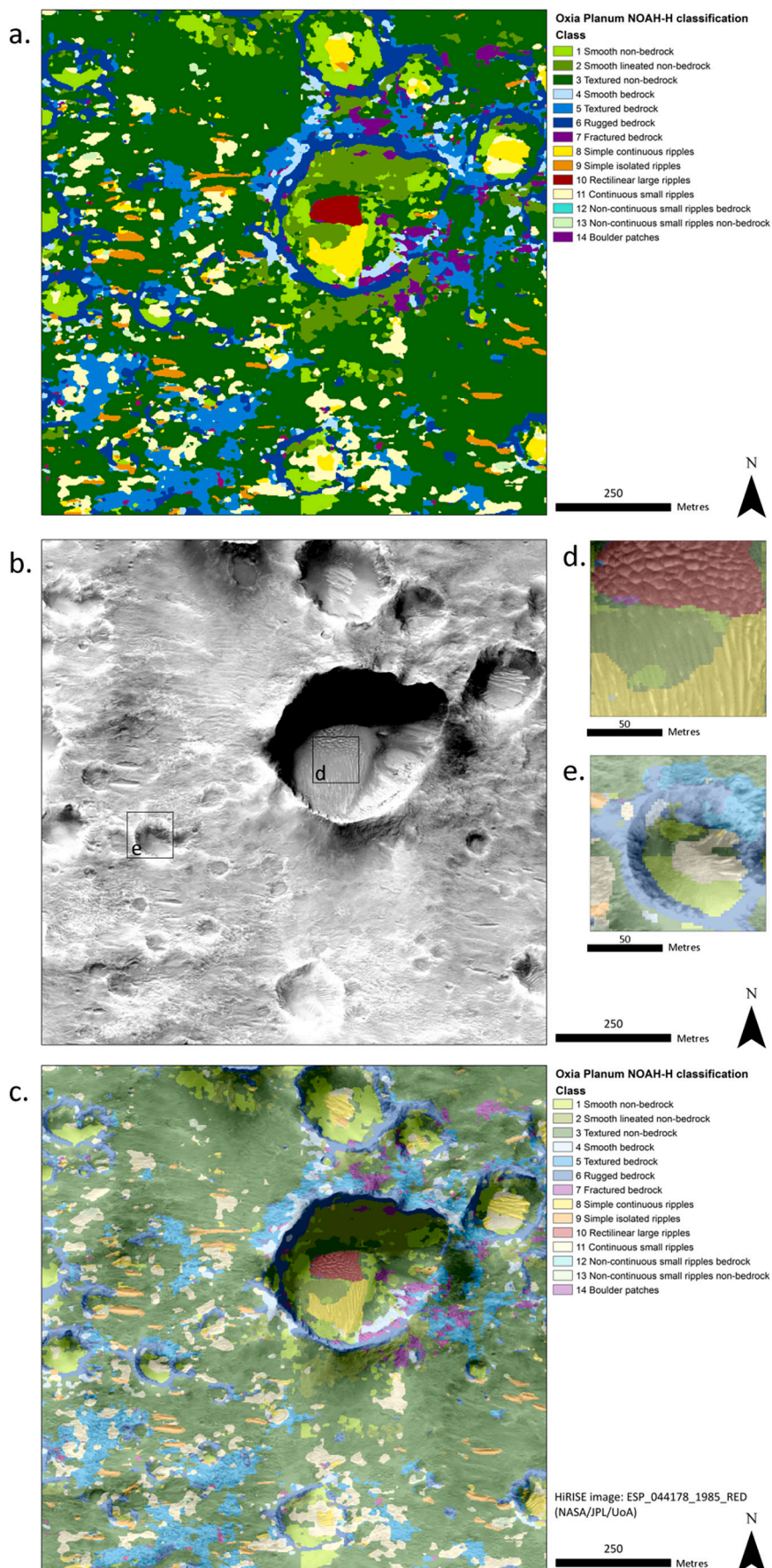


Fig. 20. Ripple filled craters in Oxa Planum. This area consists primarily of textured non-bedrock, with a few bedrock outcrops. The largest of these are the rims of the various 50–300 m diameter impact craters, which are clearly identified as rugged bedrock (dark blue). The floors of these impact craters are filled with patches of ripples, including a correctly identified region of rectilinear ripples (red). Areas of isolated and continuous ripples are also very well resolved. A few pixels around the edge of the rectilinear ripple patch are incorrectly identified as fractures, however the majority of the margins are well constrained. Fractured bedrock (magenta) is also correctly identified around the rims of the craters in the north east. a. NOAH-H output raster, b. Original HiRISE image from ESP_044178_1985. c. NOAH-H over HiRISE. d. close up of aeolian rectilinear and simple form ripples. e. close up of small crater. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3

HiRISE images surveyed and number of framelets selected for each study area.

Mawrth Vallis	Labelling Batch	Framelets	Oxia Planum	Labelling Batch	Framelets
ESP_019809_2030_RED	1	10	ESP_011937_1970_RED	1	10
ESP_026033_2020_RED	1	10	ESP_012214_1970_RED	1	10
ESP_033826_2030_RED	1	10	ESP_035303_1970_RED	1	10
ESP_035804_2025_RED	1	10	ESP_035580_1970_RED	1	10
ESP_036661_2025_RED	1	10	ESP_036714_1975_RED	1	10
ESP_036872_2025_RED	1	10	ESP_036780_1985_RED	1	10
ESP_037294_2025_RED	1	10	PSP_007019_1980_RED	1	10
ESP_037439_2020_RED	1	14	ESP_029264_1970_RED	1	10
ESP_011383_2030_RED	1	10	ESP_036925_1985_RED	1	10
ESP_014007_2030_RED	1	10	ESP_037070_1985_RED	1	10
ESP_028578_2025_RED	1	10	ESP_037347_1985_RED	1	10
ESP_032125_2025_RED	1	10	ESP_037426_1975_RED	1	10
ESP_037795_2020_RED	1	10	ESP_037558_1985_RED	1	10
ESP_038138_2020_RED	1	10	ESP_037703_1980_RED	1	10
ESP_038349_2020_RED	1	10	ESP_038125_1975_RED	1	10
ESP_038758_2025_RED	1	10	ESP_039154_1985_RED	1	10
ESP_039879_2025_RED	1	10	ESP_039299_1985_RED	1	10
ESP_040301_2025_RED	1	10	ESP_039721_1980_RED	1	10
ESP_043637_2030_RED	1	10	ESP_039932_1980_RED	1	10
ESP_043782_2025_RED	1	10	ESP_040288_1980_RED	1	10
ESP_044204_2020_RED	1	10	ESP_040433_1985_RED	1	10
ESP_044903_2025_RED	1	10	ESP_040921_1985_RED	1	10
ESP_045114_2025_RED	1	10	ESP_041066_1985_RED	1	10
ESP_045457_2025_RED	1	10	ESP_041132_1985_RED	1	10
ESP_045536_2020_RED	2	20	ESP_041211_1980_RED	1	10
ESP_045747_2030_RED	2	20	ESP_041422_1985_RED	2	20
ESP_046103_2030_RED	2	20	ESP_041989_1980_RED	2	20
ESP_046248_2020_RED	2	20	ESP_042134_1985_RED	2	20
ESP_046314_2030_RED	2	20	ESP_042556_1985_RED	2	20
ESP_046459_2025_RED	2	20	ESP_042622_1985_RED	2	20
ESP_046525_2030_RED	2	20	ESP_042701_1980_RED	2	20
ESP_046670_2025_RED	2	20	ESP_042846_1985_RED	2	20
ESP_046960_2030_RED	2	20	ESP_043057_1985_RED	2	20
ESP_047237_2025_RED	2	20	ESP_043558_1980_RED	2	20
ESP_047738_2020_RED	2	20	ESP_044178_1985_RED	2	20
ESP_047883_2025_RED	2	20	ESP_044257_1980_RED	2	10
ESP_048239_2025_RED	2	20	ESP_044323_1985_RED	2	20
ESP_049162_2030_RED	2	20	ESP_044679_1985_RED	2	20
ESP_049307_2020_RED	2	20	ESP_044811_1985_RED	2	20
ESP_049584_2030_RED	2	20	ESP_044824_1985_RED	2	20
ESP_050151_2025_RED	2	20	ESP_044890_1980_RED	2	20
ESP_050217_2030_RED	2	20	ESP_044956_1980_RED	2	20
ESP_050507_2020_RED	2	20	ESP_045101_1980_RED	2	20
ESP_050573_2025_RED	2	20	ESP_045378_1980_RED	2	20
ESP_050652_2020_RED	2	20	ESP_045523_1985_RED	2	20
ESP_050718_2025_RED	2	20	ESP_045589_1985_RED	2	20
ESP_050797_2020_RED	2	20	ESP_046156_1980_RED	2	20
ESP_050863_2030_RED	2	20	ESP_046235_1980_RED	2	20
ESP_051219_2030_RED	2	20	ESP_046301_1975_RED	2	20
ESP_051285_2025_RED	2	20	ESP_046367_1990_RED	2	20

HiRISE Images		Framelets	HiRISE Images		Framelets
Batch 1	24	244	Batch 1	25	250
Batch 2	26	520	Batch 2	25	490
Total	50	764	Total	50	740

Table 4

General dataset statistics: Total number of images for each set and the total number of Megapixels (MP). The number of MP are noted, along with an approximation of the percentage, of each image that was labelled using the DAT.

Statistic	First Run	Second Run	Final Run
Total number of framelet Images	917	1507	1507
Total number of framelet images for training	824 (~ 90%)	1414 (~ 94%)	1414 (~ 94%)
Total number of framelet 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 framelet image	~ 63.1%	~ 65.8%	~ 59.7%

Table 5

Confusion matrix showing where misclassifications occurred between the various bedrock classes.

Labelled by Expert				
Prediction	Bedrock Smooth	Bedrock Textured	Bedrock Rugged	Bedrock Fractured
Bedrock Smooth	51,074	8066	6025	4326
Bedrock Textured	13,135	391,910	262,316	61,240
Bedrock Rugged	1231	195,737	1,794,826	79,897
Bedrock Fractured	3296	86,113	98,123	1,173,215

IoU between the different classes. Are the terrains which present a very high or localised hazard reliably identified, or are they among the classes pulling down the overall average?

Rectilinear ripples exhibit the highest precision, despite the low support for this class. This perhaps suggests that the model has learnt that it is rare, and so produced few false positives. Fortunately recall is also generally high, lagging only behind large simple ripples and fractured and rugged bedrock. Rectilinear ripples are so distinct that they are rarely misclassified. This is good, since the large ripple classes present major challenges for rover navigation, so the fact that all three are easily distinguished provides confidence that the model is fit for purpose.

Other “hazardous” terrains such as boulder fields and fractured bedrock are also found with high precision, which may be due to their very distinct textures relative to the other classes. Rugged bedrock, the class which poses the greatest overall hazard, also shows strong agreement with more than 80% precision in the final run. This is encouraging from a traversability perspective.

The lowest precisions are seen for smooth and textured bedrock, while these improved in the final run, they are still not as reliably classified as the more distinct surface textures. Smooth bedrock is one of the “safest” classes to traverse, so any confusion with this is unfortunate. However, textured bedrock does not have high relief and so is generally safe. It might pose localised hazards but these would have to be identified *in situ* regardless of the model result. We thus do not consider this to be a dangerous form of confusion.

Lower precisions were also found for the small ripple classes. The model struggles to identify whether small ripples are continuous, and whether they are found over a bedrock or non-bedrock substrate. While none of these are the most dangerous classes to confuse, lower agreement here still limits the applicability of the model. Substrate is relevant to how much grip the rover might have when crossing a ripple field. However, not being able to determine this is not as critical as failing to identify a region which would be entirely impassable. Grouping the classes reduces variation in precision. Small ripples remain less reliably classified, but they are still correct in excess of 70% of cases in the final run. Section 6 will show that lower precision for this class at the pixel scale did not adversely affect the usability of the data at a landscape scale. These lower precisions are thus not ideal, but do not make the classified raster unfit for purpose.

The recall results show a similar pattern, especially when classes are grouped. However, there are some minor differences. For many classes recall is slightly lower than precision. Just because the model is able to accurately label the examples that it has found, does not necessarily mean that all possible examples have been detected. This has important implications for the traversability assessment. If areas of low hazard are incorrectly classified as being dangerous, then this is likely to become apparent when a human operator inspects the area in order to decide how best to navigate around the hazard. A decision can then be made as to whether the model was correct in its classification. However, hazardous terrains which are classified as safe by the model may not be inspected with the same care. Several classes; smooth bedrock and large continuous ripples have higher recall than precision. Small ripples over bedrock exhibit the greatest disparity. This indicates that most true examples of this terrain have been identified, but that many other landforms are also incorrectly labelled as this class.

Recall results are high enough that the model remains fit for purpose, especially since the results are intended to augment a human workflow rather than replace it. However since recall is never perfect, the model should not be considered a sole arbiter of hazard, and should always be interpreted critically.

The increase in agreement across all metrics when classes are combined into interpretive groups is also encouraging since this demonstrates that the majority of confusion is occurring within groups, rather than between them. While the model may struggle to distinguish between different ripple forms, or different surface textures, it is rarely

confusing a field of ripples with a bedrock surface, or a boulder field with ripples.

Non-Bedrock exhibits the least confusion with other groups, scoring a 96% prediction, with negligible amounts of confusion dispersed evenly among the other groups.

Similarly, Bedrock scores well at 94% with the majority of the confusion caused by the Non-Bedrock Group. Large Ripples score 92% with all of the confusion shared with Small Ripples. However, the confusion between Large and Small Ripples is greater when predicting Small Ripples with over 20% of the confusion caused by this class. Boulder fields score reasonably well with 85% confidence. Bedrock is the main source of the confusion in this case. The groups, in and of themselves are not sufficient for traversability assessment, since there are variations in hazard within each (for example whether ripples overlie bedrock or not, the continuity of ripple fields, and the roughness of bedrock surfaces.) However they demonstrate that the model does well at predicting the thematic character of the site.

The majority of confusion occurs between classes such as rugged and fractured bedrock, which frequently form a continuum. When trying to predict Fractured Bedrock the most confusion (7.2%) comes from Rugged Bedrock although Textured Bedrock also causes a degree of confusion albeit to a lesser extent. When predicting Rugged Bedrock this pattern is reversed with the main confusion coming from textured bedrock (9.2%). Both rugged and fractured terrains are potentially hazardous. Rugged terrain will always limit rover navigation, while the extent to which fractured terrain will be navigable might depend on the relief of the fractures and the extent to which they are infilled. Both would be best to avoid. The classification as potentially hazardous invites manual inspection which can quickly identify which areas might be situationally traversable.

5.6. Output examples

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 ellipses. In total, 12 images were selected for Oxia Planum and 18 for Mawrth Vallis. These images were those which provided the most central coverage of the 1-sigma landing ellipses and so were most critical for the landing site selection exercise. Post-processing of the NOAH-H output included down-sampling to 2 m/pixel and conversion to colour classes for analysis.

Each of the classes was given a unique combination of RGB values in order to create masks with the classification. This allowed the output to be displayed conveniently as a different colour for each class in image-viewing software, and then to be easily converted into a single band product in a GIS. This allowed us to formally manipulate the data and to overlay the NOAH-H output onto the original HiRISE images for inspection. Figs. 12–14 show a “worked example” of a small, representative HiRISE image. Fig. 12 shows the colour codes chosen for all classes along with an example HiRISE image showing the distribution of labelling framelets, and the polygons with which they were labelled.

The examples in Fig. 13 show the output of the model for an entire HiRISE image. This demonstrates how the model is able to learn from a small amount of data to segment the whole region with high accuracy, producing a result with strong agreement to the manually labelled set. It immediately provides a huge amount of classified data, which would have been laborious to acquire manually.

Fig. 14 shows examples from the validation dataset for this image, to compare what was manually labelled, and what the system classified for the First and Final Runs. In the “Final Run” the system can very accurately predict the labels assigned by the expert geomorphologists. It is also able to reliably classify the remaining un-labelled pixels.

There are some aspects of the campaign presented here which mask the underlying performance of the algorithm versus the expert, which is not fully captured in the mean IoU. This is because the expert geologists have labelled portions of frames individually without labelling the

neighbouring frames; this means that the experts had the contextual information around the tile, but the DL trained model did not. A like for like comparison would require the geologists to label all of the pixels in the frames and in addition also label the neighbouring frames.

6. Reliability and applicability

6.1. Landscape level reliability

In order to assess whether the classification is fit for purpose we must consider the geomorphological context of the results in Section 5. Considering accuracy not just on a pixel by pixel basis, but also on a “landscape scale”. The impact which errors will have on the usability of the data by a human operator depends greatly on how the misclassified pixels are distributed and between which classes the confusion occurs.

In order for a landform to be observable in a raster image it must cover multiple pixels. Thus the spatial-scale at which variations in surface geomorphology can be recognised using HiRISE is $>1\text{--}2\text{ m}$ or $4\text{--}8$ pixels. Misclassified pixels reduce the accuracy of the model as a whole. However, their actual impact on the ease with which the output can be interpreted is highly situational and depends upon the tasks to which the data is put.

Pixel scale accuracy is important when performing additional geospatial operations or manipulating the output raster using GIS. If the classification were to be fed directly into a hazard avoidance system then $\sim 74\%$ agreement at the pixel scale might not be sufficient. However, for other purposes, such as locating terrains of a specific type or preparing geomorphological maps, it is more important that the broad trends are correct on the $4\text{--}8$ pixel scale. Since this model is not intended to serve as a sole arbiter of hazard or geological origin it is most important that the results be easy for a human operator to interpret, so that they have the information needed to reliably make those next steps. Frequently this does not require the specific pixel-point pairing to be correct at every possible point, but rather that it is correct on average over the $4\text{--}8$ pixel scale.

Fortunately, this is what we see when examining the classified raster. Isolated incorrect pixels were often found in areas which were otherwise classified correctly. Coherent patches of misclassified terrain were less common (though they do occur). Erroneous pixels result from a variety of factors such as subtle variations in the surface texture at the metre scale, or a small “inclusion” of one terrain type within an area predominantly classified as another. These factors are discussed in more detail in Section 7.

A large patch of coherently misclassified terrain would present a major challenge to interpretation of the raster by a human, whereas an equivalent number of inaccurate pixels, scattered randomly throughout a large area of otherwise correctly classified terrain would have little impact upon a human interpretation of the results. This means that, when creating summary products, landscape level accuracy can be improved by down-sampling the data using an appropriate averaging scheme. This removes isolated misclassified pixels and provides a product which is more accurate than the pixel scale IoU would suggest. The result is much closer to the sort of summary product produced by a human team, which would not attempt to classify every pixel scale variation.

6.2. Qualitative assessment of representative examples

We identified classes with lower IoU, which were frequently confused. By examining areas where such confusion occurred the geomorphology team were able to identify landscape characteristics which frequently result in misclassification. Consequently, we will be able to account for these trends in future work, using this dataset for science tasks.

Figs. 15–20 show $\sim 1 \times 1\text{ km}$ square examples of NOAH-H output from the two study areas. Each figure consists of 3 images; the classified

raster produced by the NOAH-H model, the original HiRISE image from which it was produced, and a translucent NOAH-H layer, overlain on the HiRISE image it classifies. Smaller insets provide a close-up view of key features. A slightly different colour scheme has been used for these figures, to better highlight the relationships between similar classes and to provide better contrast when draped over HiRISE data. The colour scheme has been applied consistently across Figs. 15–20. These examples were chosen because they are representative of the study areas as a whole rather than being particularly good or bad cases. Each image highlights one or more phenomena which affects the interpretability of the classified raster, and which should be taken into consideration when using the data.

NOAH-H was found to be very good at identifying the key features of the HiRISE image, even if it did not always digitise them with 100% reliability. Large discrete features such as large ripples and the rims of craters are generally reliably identified, matching the high IoU for these classes. The “landscape level” character of the site is very well expressed by the NOAH-H results. However, many of the specific boundaries are open to interpretation. Consequently, while the classifier provides a valuable tool for further manual investigation, it is not able to produce a complete map by itself.

Ripple forms were segmented most accurately of any of the groups. The edges of large ripples are very well defined, and patches are usually found reliably. There are, however, some peculiarities in how different ripple forms are classified. Fig. 15 shows a case where ends of ripples protruding from the main field have been classified as isolated (orange), while the rest of the ripple is continuous (yellow).

A human mapper would be unlikely to assign two parts of the same feature different classifications. However, the two ends of these ripples do conform to different class descriptions, and the transition between the two is well defined. The model lacks the contextual understanding that a human would intuitively use to decide what constitutes part of the same feature. Splitting of ripples is also seen in Fig. 15, and occurs frequently across both study areas. While this is perhaps not how a human would digitise this landscape, it is not per se wrong and would not adversely affect traversability assessment. It should be noted that these peculiarities only occur in the discrimination between different ripple classes. When the broader group of ripples are considered as a whole, they are distinguished from other terrain types very reliably.

More subtle variations such as transitions between different surface classes are not always as precisely segmented. There are also some artefacts, where small variations within a terrain were incorrectly classified as being examples of a different class. In any real landscape there is usually a dominant terrain, with small areas of other classes interspersed or overlain. In many cases the NOAH-H system was able to outline these smaller patches, but the boundaries were not always exact.

While this limits the reliability of the data, it does reflect reality, since the transition from one surface to another is often gradational. The model is designed to segment areas exactly, and so has no way to deal with “fuzzy” boundaries. For example Figs. 17 and 18 both show cases where fracture networks extend beyond the areas classified in magenta, but become too subdued to be recognised by the model.

Fractures grade into hummocky ground which has been classified as textured or rugged bedrock (mid & dark blue). A human is likely to give this terrain the “benefit of the doubt” since they can see that it adjoins a fracture pattern, and recognise the troughs as degraded fractures. Without that context, the AI decides where the boundary should lie using purely morphometric criteria, excluding areas which could arguably belong to either class. This highlights an inescapable issue that even humans struggle with – how to classify an area that has characteristics of more than one class (i.e., a fractured, rugged bedrock area).

In most cases where the exact boundaries could be disputed, the majority of the feature’s area is found to be correctly classified, leading to the generally high IoU reported in Section 5. The model’s pattern recognition is good, but it cannot make an interpretive leap based on the context of a feature. This results in what a human would consider an

incomplete classification.

Boulder fields present an interesting challenge for the model, since they are discontinuous features, which overlie terrain of other classes; there are usually relatively large spaces between the blocks. This seems to make it hard for the model to digitise the full extent of the patch. An example can be seen in Fig. 15. Small patches of boulders have been identified along the rim of the southern crater. However, many areas of boulder strewn ground are grouped in with the rugged or textured bedrock (light & dark blue) which they overlie. The purple patches give an indication of where boulders are found, but do not segment their full extent. The classified raster meets the objective of providing a useful starting point for a human mapper but falls short of being able to digitise the terrain itself.

Full digitisation of the boulder field is challenging, since there is no sharp divide between one class and another. Rather, as boulder spatial density decreases beyond a certain threshold, the terrain is no longer appropriately classified as being a boulder field, and the designation of the underlying terrain type should instead be adopted. A geomorphologist digitising this landscape would have to decide where this threshold lies, and the same is true for the AI. The model was never specifically programmed to follow the criterion of 10 blocks per 10 m² (the ‘rule of thumb’ used to define boulder patches during the labelling exercise). Rather it learnt what a boulder patch was by example. The same is true of small isolated ripples in the 0.5–2 m range. Many areas of these are correctly identified, but not all patches can be distinguished from the background.

The extensive boulder field in Fig. 16 provides another example. Although most of the areas with the densest boulder coverage are identified (purple), there are still dense patches of blocks within the areas identified as bedrock (blue), or non-bedrock (green). There appear to be more unclassified blocks in the rugged bedrock areas, perhaps suggesting that it is harder for the algorithm to distinguish boulders from rough terrain than from the generally smoother non-bedrock. However it could also indicate that the boulder fields are more continuous over the bedrock surfaces, while the non-bedrock areas have more self-contained patches of boulders.

A more precise delineation could be attained by training a model with several different classes of “boulder field”, each reflecting a different substrate or spatial density. This would be worthwhile in an investigation focused on boulder distributions, but was not feasible in this more general case.

In Fig. 19 there are a few areas where patches of red indicate the presence of rectilinear ripples; however no example of this morphology is apparent in the HiRISE image. Rather, these areas contain a bedrock ridge, running perpendicular to the crests of the small bedforms, which gives them their rectilinear appearance. The ridge is more correctly identified as rugged bedrock (blue) in other parts of the image, where it is not overlain by ripples. To the east side of the image, the ridge feature is not very distinct, and is generally classified as textured non-bedrock (green). This shows how the system can be “fooled” by unusual superpositions of unrelated landforms, for which it has not been trained.

Fig. 20 shows an interesting result of the classification. The extensive shadowed region to the north of the largest crater has been classified as smooth lineated non-bedrock (pale green), with some areas of textured non-bedrock (dark green). The surface features are not apparent to a human surveyor, since they are impossible to distinguish with the default grey-scale stretch of the HiRISE image. They only show up when the stretch is manipulated using post processing. However, the model has been able to distinguish these subtle features correctly since the lineated pattern becomes distinct when the convolution filter examines the image through a small moving window.

6.3. Impact on traversability

The most important classes to be distinguished for this test case were those which directly impacted landing site selection. This involved

considerations such as the density of aeolian and clastic cover, and the broad distribution of fractured and rugged bedrock terrains. Large ripples present a larger overall hazard than smaller ones. Larger ripples will never be traversable, whereas smaller ones might in some cases. This would depend on situational factors which would have to be assessed in situ.

Other determinations were less critical. All non-bedrock terrains represent fairly uncertain levels of hazard, so while it is important to be able to distinguish them from bedrock, identifying the specific sub classes is less critical. Lineated non-bedrock was almost always found on steep slopes, such as crater walls which would be avoided based on slope considerations. In this case the interpretive group, with its higher overall IoU, was sufficient for the task of hazard identification, whereas this was not the case for other features.

Table 5 shows the confusion matrix for the bedrock classes. This is an important group, since it contains both rugged and fractured bedrock, the two surface classes where the impact on traversability is inarguably high. Like the non-bedrock classes these surface textures form a broad continuum. However in this case it is vital that various members of this series be distinguishable. The most confusion occurs between rugged and textured bedrocks. However, fractured and rugged bedrock were generally very well differentiated from smooth bedrock, the least hazardous terrain within this group. It is only in the case of textured bedrock, the most “catch all” of the series that the model confuses a significantly large proportion of identifications. These are thus not especially “dangerous” mistakes.

Confusion between rugged and fractured bedrocks could be of concern, since fractured terrains could be traversable depending on the exact arrangement of troughs and the extent to which they are infilled. However, since this is situational, and subject to localised hazard, further analysis by the human operator would be needed before such a route could be planned and so an error would be noticed at this stage.

The only “dangerous” mistakes are seen in the cases of small ripples, where the model cannot always distinguish between bedrock and non-bedrock substrates. Small ripples over bedrock are more frequently confused with continuous small ripples, whereas small ripples over non-bedrock are correctly detected in a higher proportion of cases.

These results show that the model is good enough for its intended purpose. It cannot provide a definitive answer on whether a terrain will be hazardous, since this would require comparison with topographic data, and consideration of in situ soil type data, and the engineering constraints of the rover. However, the NOAH-H algorithm can indicate areas where certain textures are most common, and give a human operator a head start in studying those features in more detail.

Both the individual classes, and the broader interpretive groups serve a useful purpose. The groups can indicate the broad character of the site with very high precision and recall. The operator can then refer to the full class distribution in order to examine the intra-group variability in more detail. This allows them to make observations of how and why certain areas have been correctly or incorrectly classified, and build a more in-depth understanding of the area. In this respect the two tiers of the hierarchical classification scheme are complementary. Neither provides a perfect solution on its own, but when considered together a good understanding of a classified site can be attained.

7. Limitations and considerations

Several sources of uncertainty were identified during the labelling process. Every attempt was made to limit their impact on the study as outlined below. Labelling also highlighted several areas in which the classification system could be improved upon, most significantly combining easily confused classes, in cases where the distinction between them is not vital for the study.

7.1. Variability between and within classes

Although a set of discrete ontologies were defined, many of the textures formed part of a continuum. This is particularly true of the surface classes. Some regions had attributes of both rugged and textured bedrock, thus forming an intermediate step between the two ontologies. Without using a prohibitively large set of classes it was not possible to treat every possible variation as a discrete class. Every effort was made to remain as consistent as possible when deciding which ontology a transitional terrain represented and they were left blank when classifying them proved impossible. Such determinations are inherently subjective, so a different geomorphologist might consider the cut-off between classes to lie in a slightly different place. In order to reduce subjectivity a single member of the science team conducted the labelling, and remained as consistent as possible.

Attempting to classify all possible surfaces into one of fourteen classes presented various challenges. It was inevitable that there would be variation within each class. This can be seen in the type examples presented in [Section 3.1](#), subtle variations in morphology and texture are present in almost all classes. Attempting to split continuous variations into discrete categories will always involve some degree of subjectivity.

7.2. Transitions and inclusions

Transitions between regions of different classes were often difficult to define. In cases where the exact boundary was unclear, the boundary region was not digitised. The two 'blocks' where the surface type was identified with high confidence were labelled, and the area in between was left blank. While this approach made the labelling work practical, it did have implications for the classification. Since the NOAH-H algorithm classifies every part of an image, and does not use a background class, it was not able to ignore transition regions which did not conform to a clear ontology. Examples of the effect this had on classification are illustrated in [Section 6.2](#).

In some cases, mixed patches occurred, where small regions of one terrain type were found in very close proximity to others, or a large area of one terrain type contained smaller regions of another. In general, these "inclusions" were avoided during labelling. However, in some cases the presence of an inclusion was itself a genuine feature. An example is textured bedrock. The texture often arises from the fact that undulating bedrock has small pockets of non-bedrock material within it. In such instances, care was taken during labelling to avoid any area which was large enough to be digitised as a separate non-bedrock patch, while smaller patches which formed part of the characteristic texture were included. This approach was applied consistently, and does seem to be reflected in the way the algorithm ultimately classified similar terrains.

Many discontinuous features were too small to be digitised in isolation. In the case of boulder fields, it would have been prohibitively time consuming to draw around each boulder. Thus, the entire field was outlined, including whatever material the boulders overlay. This does appear to have affected the model's classification, in some cases areas classified as boulder patches include a "buffer" around the outside of the feature, where it overlaps non boulder covered terrain. It seems likely that the model has learnt that boulders must be surrounded by a buffer of open space. It has then applied this to the edges of the patch, not just the area between boulders.

Sometimes a landform was defined by its surroundings. For example, the distinction between the two classes of non-continuous small ripples relies on whether they overlie bedrock, or non-bedrock material. In this case the surrounding material was included in the patch, rather than digitising each small ripple individually, as was done with larger ripples. It might have been useful to include a similar distinction for other distributed features such as boulders, which can also overlie various surface classes. However, this distinction was not as relevant to the

present study.

An understanding of the surface type which ripples overlie is potentially important for traversability, as a Rover might be able to gain traction when several of its wheels are on the bedrock areas surrounding a sandy ripple, but not if the ripple was surrounded by looser regolith. In contrast, a boulder patch remains a navigational hazard irrespective of the underlying terrain. For other studies aimed at determining the specific geomorphology, as opposed to the traversability, of the surface, such considerations will be more relevant.

7.3. Interactions

In some cases an interaction between two adjacent textures influenced their classification. For example, when a single large ripple was found within a patch of continuous smaller ones there were two possible classifications. It could have been defined as an isolated ripple, since it was not adjacent to any other large ripples. Conversely, it could have been defined as an example of continuous large ripples, since it was surrounded by other ripples, albeit of a smaller size. It was decided that the second option provided the best description, and this was consistently applied throughout the labelling process. In these cases, the small and large ripples were not labelled as a single feature, rather the large ripples were digitised separately, as would be the case for an isolated ripple. The smaller ripples were then digitised alongside them.

Other transitions, such as the boundaries between simple and rectilinear ripples, or between continuous and non-continuous ripples were treated in a similar manner, and when in doubt, a buffer of unlabelled terrain was left between the two patches.

In some cases, it was not possible to make a determination as to whether a patch of ground consisted of bedrock, or non-bedrock material. This determination was harder to make when only the small framelet and its context image were available. Features such as rough or fractured ground, which might indicate that the terrain is clearly 'bedrock', can easily fall outside the field of view of a small framelet. Were a larger image available, it would have been possible to identify indicators of bedrock in other areas, and follow the contact between the two terrains into the ambiguous area to make a determination. If that contact fell outside the field of view then this was not possible, and other indicators such as consistent albedo or similar texture had to be used to determine that two blocks of terrain were part of the same class.

Finally, large patches of continuous ripples were often found to extend beyond the edge of the framelet context area. In these cases, if labelling only the adjacent framelet, then the ends of the ripples would appear to be 'isolated ripples', despite the fact that they were actually continuous, with other parts of the ripples touching to form a continuous patch. However, because this information exists beyond the edge of the framelet and its context image, it could not inform the labelling.

7.4. Scale

Intra-class variation is particularly important when considering scale. A patch of rugged terrain on a metre scale can look very different to rugged ground on the decametre scale. The latter often consists of rugged scarps and ridges, while smaller scale rugged areas are much flatter. When examined independently of scale both have similarities, but it is not a given that planetary surfaces have a "fractal" appearance. Thus, when seen together within the same framelet, these terrains could be considered different textures.

This is true of many classes where the same feature occurs on multiple scales, including aeolian bedforms. During the labelling phase it became clear that it would have been useful to retain an intermediate size category when defining the scale of ripples. Placing the cut off between large and small at ~5 m across (as measured perpendicular to the ridge crest) meant that the "large ripples" categories included a very wide range of scales. However, this was intended to represent features that were likely impassable to a rover, compared to features that could

plausibly be traversed (e.g. Balme et al., 2018). Alternatively, since the size of a digitised feature is easy to measure after the fact, the distinction between small and large ripples might be unnecessary.

7.5. Size of the training dataset

The reliability of the NOAH-H results and their applicability depend greatly upon the quality and representativeness of the training data. The NOAH-H dataset is very small by machine learning standards with only ~1500 training framelets. In contrast one of the most popular datasets for applications like object detection, segmentation, and captioning, COCO (Common Objects in Context; Lin et al., 2014) contains 124 k images. The factors that determine the perfect dataset size are mainly: (i) the classifier architecture in question - in particular the parameters of the model, (ii) the number of features considered, and (iii) the statistical characteristics of the data - the variation in the samples for each class.

The small size of the NOAH-H dataset proved challenging. Since it contained only 1504 images for 14 classes, meaning that several classes had very few examples in total. DL approaches are often said to require enormous amounts of data to work well, but our results indicate that it is still possible to obtain good performances with limited data. We employed techniques to prevent overfitting such as dropout, weight decay, data augmentation, pre-training, and parameter sharing. This approach enabled us to train very large models with millions of parameters on this dataset. However, additional data is still needed in order to improve the generalisation of the model and achieve more accurate results.

Though small by machine learning standards, our dataset is significant in the context of planetary geomorphology. The labelled frames are representative of the area that they sample, and were digitised with high precision. The labelling work carried out in the NOAH-H activity and the parent NOAH project represents the largest collection of its kind for an application such as this. Labelling images of this type is extremely labour intensive, especially in cases where specialist knowledge is needed. The number of experts available globally for such a niche task as classifying planetary surfaces is extremely small and their time is often over-subscribed. Consequently, the development of this dataset has been a major achievement for this work. In machine learning terms this is of course a relatively limited dataset but in the context of space science research it is significant and has produced meaningful results.

8. Discussion

It is important to note that this was a prototyping activity. More work is required to produce a fully mature system. The results are encouraging and with additional training the overall accuracy of the system is expected to improve to an even finer level of granularity. Expert supervision is still needed to correct any possible misclassification error in the short to medium term.

The key question to consider is whether the automatic classification is good enough for the intended purpose. This must be considered both in absolute terms, and in comparison to the quality of human mapping efforts. Manual mapping would not typically exhibit perfect precision and recall on a pixel scale, were it to be compared to ground truth. Rather an appropriate mapping scale is chosen for the task, and geomorphological units are repeatably identified at this scale, with an aim of showing the characteristics of a site. While 100% IoU will always be desirable, it is not typically attainable in a manual mapping campaign, and is rarely assessed.

Qualitative study by experts on the geomorphology team has confirmed that, on a landscape level, the quality of the classified rasters is comparable to what we would expect of a manually classified area (see discussion in Section 6). The model is not reliable enough to be used as a sole arbiter of hazard, but as discussed above, this was never the intention. Rather it forms a useful tool to assist ongoing human efforts. In this respect it more than meets the requirements of the study. The

agreement between the model predictions and the validation data is good enough for purpose. We thus conclude that the classified rasters are suitable for use in the ongoing characterisation of Oxia Planum, and as part of strategic planning as the ExoMars mission advances.

Summary products showing the abundance of different potentially hazardous classes can be produced, and down-sampling can be used to reduce pixel level inaccuracies. The HiRISE resolution is high enough that this will not have an adverse effect on the usefulness of the data. Variations in precision, recall, and IoU should be kept in mind when judging the reliability of identifications of specific classes, but none are far enough out of bounds to be disregarded in future use cases.

The ontological classes selected for this work have proven capable of identifying areas which would be safe to traverse, and regions dominated by aeolian bedforms, or loose non-bedrock textures (as shown in Section 6), which provide significant traversability risks. The results closely match the ongoing expert mapping of the sites by the landing site selection team. NOAH-H revealed areas where aeolian bedforms were more or less common, and was able to discriminate well between the most important bedform classes. A higher proportion of bedrock was found in Oxia than Mawrth. Areas of fractured bedrock were identified, which correlated with clay-bearing units (Quantin-Nataf et al., 2019).

The results of this study also have applicability beyond the specific task of traversability assessment. The ontological classes could easily be altered to describe the textural geomorphology of other regions, and many classes are already directly transferable to other tasks (e.g., the aeolian bedforms and boulder patches). The reliability with which aeolian bedforms of different morphologies are detected has great potential to identify these features in other HiRISE images, and potentially develop a global catalogue of transverse aeolian ridges.

We advise that both the full class list, and the interpretive groups be considered in parallel when using this data to make determinations. The groups provide higher IoU, and summarise the landscape in interpretive terms. However, the descriptive classes are required to fully judge the traversability of the landscape.

We conclude that true geomorphological mapping would be challenging to achieve with this method due to the degree of interpretation required. The human ability to link observations of textures and shapes, and designate them as “landforms” is based not only upon observations, but also knowledge of the processes that created or altered those landforms. This is developed through personal experience investigating similar examples on Earth. Replicating such expertise using machine learning would be an extremely daunting task, although perhaps approachable with a sufficient volume and diversity of training data. In many ways a descriptive model is more powerful than an interpretive one, since the more general surface classes defined herein can be applied to a wider variety of tasks than stricter geomorphological class labels.

The strong agreement exhibited in these results combined with the very high processing speed allows a scientist to use the NOAH-H classified data to assist their interpretation without having to perform time-consuming mapping or manual image classification. However, it is important that ontological classes are well-defined, and that sufficient training data volumes are labelled – which is itself a time consuming task. There therefore exists a trade-off: studies with low data volume (a few HiRISE images) are perhaps performed more quickly using manual digitisation, but if a large data volume is to be analysed, the time investment in labelling is worthwhile, as the automated method will be quicker in the end.

The initial labelling campaign took approximately 270 h to complete, producing a total of 236 MP of labelled data. The model has since been run on ~160 HiRISE images with a combined area on the order of $\sim 2 \times 10^{11}$ pixels. A human geomorphologist working round the clock would have taken 26 years (2×10^5 hours) to complete the work, assuming that the full area could be mapped at the same rate as the individual framelets. Using the NOAH-H system, complete images can be fully classified at the pixel level in a matter of hours. Even factoring in months of model development work in addition to the time taken for

labelling and image processing, the Machine Learning approach fundamentally changes the scope of what a study of this sort can expect to achieve.

8.1. Further work

There are various ways in which this work can be built upon. The present work has been limited in scope since the focus was to classify the candidate ExoMars landing sites. However the same method could be applied to other areas of Mars, and other questions both scientific and operational. The framework presented in table one can be used to build compatible classification systems, tailored to landscapes with a different variety of terrains to Arabia Terra. With enough training data it could be possible to train a more generic model, which could be applied to any region on Mars, however the more practical approach would likely be to continue to focus on more limited areas, using a smaller set of ontologies tailored to those locations.

If it were possible to collect and label more training data then the overall output of the model could be improved. This would also open up the possibility of other lines of investigation such as training a model for each study area independently and then testing how well the results transfer to the other site.

One interesting comparison, which was not in scope for the present project, would be to compare the results of the machine learning approach to other algorithms and classification techniques. More useful still would be a direct comparison to manual mapping of the area, although by its very nature this would have to be limited to a much smaller area than any automated technique.

It would be interesting to apply this technique to the landing sites of past Mars rovers, where ground truth data could be used to interpret the descriptive classes identified from RS data. This would allow more detailed traversability assessments to be conducted, and answer questions about the geological origin of the features which the model has been trained to detect. It will of course become possible to address these questions at the Oxia Planum study area once the *Rosalind Franklin* rover begins to return in situ data.

9. Conclusions

The results of the study are encouraging. The NOAH-H model exhibits good agreement with the validation data, and produces reliable results, which assist in locating hazardous terrains, and inform our understanding of the candidate landing sites. The final run of the model produced a mean IoU of 74.15% for the full list of surface-type ontologies and 92.33% when considering groups of similar surface-types. The higher accuracy for groups of classes demonstrates the need for larger training datasets in future.

Some classes were more reliably classified than others. Distinct features such as large scale aeolian bedforms, rugged and fractured bedrock surfaces, and boulder patches were generally identified with high precision and recall. The model performed less well for small scale ripples, non-bedrock, and smoother bedrock classes. Many of these ontologies describe a continuous morphological spectrum. Hence, dividing them clearly into discrete ontological classes proved challenging, both during labelling and for the model. In this respect, the model mirrors the human difficulty in classifying a continuum of forms into discrete classes.

When examined in detail, pixel scale inaccuracies in the classified rasters were found to not substantially detract from their usefulness as a tool to describe landscape level trends in the distribution and prevalence of surface textures. They provide surface texture maps that are representative of the site. This provides a useful first step in more detailed mapping activities, or for delineating terrains that would be a hazard to rover surface operations. We thus conclude that the data is fit for purpose. The NOAH-H output is already being used to identify terrains of interest, such as areas with aeolian bedforms (Favaro et al., 2021). It will continue to prove useful as characterisation of the Oxia Planum landing site continues.

Declaration of Competing Interest

None.

Acknowledgements

AMB and MRB acknowledge funding from the UK Science and Technology Facilities Council (STFC; grant ST/T000228/1) and by a European Space Agency contract. MW, SK, and DP acknowledge funding from a European Space Agency contract (ref = 4000118843/16/NL/LvH – Novelty or Anomaly Hunter (NOAH)). The authors have no conflicts of interest to declare.

References

- Allender, E., Stepinski, T.F., 2017. Automatic, exploratory mineralogical mapping of CRISM imagery using summary product signatures. *Icarus* 281, 151–161. <https://doi.org/10.1016/j.icarus.2016.08.022>.
- Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. <https://doi.org/10.1109/TPAMI.2016.2644615>.
- Balme, M., Berman, D.C., Bourke, M.C., Zimbelman, J.R., 2008. Transverse Aeolian Ridges (TARs) on Mars. *Geomorphology* 101, 703–720. <https://doi.org/10.1016/j.geomorph.2008.03.011>.
- Balme, M., Robson, E., Barnes, R., Butcher, F., Fawdon, P., Huber, B., Ortnier, T., Paar, G., Traxler, C., Bridges, J., Gupta, S., Vago, J.L., 2018. Surface-based 3D measurements of small aeolian bedforms on Mars and implications for estimating ExoMars rover traversability hazards. *Planet. Space Sci.* 153, 39–53. <https://doi.org/10.1016/j.pss.2017.12.008>.
- Bandeira, L., Marques, J.S., Saraiva, J., Pina, P., 2011. Automated detection of Martian dune fields. *IEEE Geosci. Remote Sens. Lett.* 8, 626–630.
- Bandeira, L., Ding, W., Stepinski, T.F., 2012. Detection of sub-kilometer craters in high resolution planetary images using shape and texture features. *Adv. Sp. Res. Sp. Res.* 49, 64–74. <https://doi.org/10.1016/j.asr.2011.08.021>.
- Bandeira, L., Marques, J.S., Saraiva, J., Pina, P., 2013. Advances in automated detection of sand dunes on Mars. *Earth Surf. Process. Landf.* 38, 275–283. <https://doi.org/10.1002/esp.3323>.
- Cadogan, P.H., 2020. Automated precision counting of very small craters at lunar landing sites. *Icarus* 348, 113822. <https://doi.org/10.1016/j.icarus.2020.113822>.
- Chen, L., 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv e-prints*, 1706.05587.
- Chen, L., Papandreou, G., Member, S., Kokkinos, I., Murphy, K., Yuille, A.L., 2018a. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Recognit. Mach. Intell.* 40, 834–848.
- Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), *Computer Vision – ECCV 2018*. ECCV 2018. Lecture Notes in Computer Science, 11211. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_49.
- Dundar, M., Ehlmann, B.L., 2016. Rare jarosite detection in crism imagery by non-parametric Bayesian clustering. In: 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1–5. <https://doi.org/10.1109/WHISPERS.2016.8071747>.
- Favaro, E.A., Balme, M.R., Davis, J., Grindrod, P.M., Fawdon, P., Barrett, A.M., Lewis, S.R., 2021. The Aeolian environment of the landing site for the ExoMars Rosalind Franklin rover in oxia planum. *Mars. J. Geophys. Res.* 126 <https://doi.org/10.1029/2020JE006723>.
- Foroutan, M., Zimbelman, J.R., 2017. Semi-automatic mapping of linear-trending bedforms using ‘self-organizing maps’ algorithm. *Geomorphology* 293, 156–166. <https://doi.org/10.1016/j.geomorph.2017.05.016>.
- Ghosh, S., Stepinski, T.F., Vilalta, R., 2010. Automatic annotation of planetary surfaces with geomorphic labels. *IEEE Trans. Geosci. Remote Sens.* 48, 175–185. <https://doi.org/10.1109/TGRS.2009.2027113>.
- Harris, E., Martin, D.J.P., 2020. Traversability study of Jezero crater using comparative orbital and ground-based image analysis of MER and MSL rover traverses. In: 51st Lunar and Planetary Science Conference, p. 1.
- Hartmann, W.K., Neukum, G., 2001. Cratering chronology and the evolution of mars. *Space Sci. Rev.* 96, 165–194.
- Jasiewicz, J., Stepinski, T.F., 2012. Global geomorphometric map of mars. In: *Lunar and Planetary Science Conference*, 43, pp. 1347–1348.
- Karachalios, S., Woods, M., Schwenger, S., Joudrier, L., 2019. Novelty or Anomaly Hunter: Towards Flight Ready Autonomous Science Using State of the Art Machine & Deep Learning. *ASTRA*, p. 2019.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444. <https://doi.org/10.1038/nature14539>.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, pp. 740–755.

- Lin, H., Tarnas, J.D., Mustard, J.F., Zhang, X., Wu, X., 2018. Dynamic Aperture Target Transformation (DAIT): a novel and valuable method for mineral detection on Mars. In: 49th Lunar and Planetary Science Conference 2018, p. 2083.
- Liu, C., Chen, L., Schroff, F., Adam, H., Hua, W., Yuille, A., Fei-fei, L., 2019. Auto-deeplab: hierarchical neural architecture search for semantic image segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 82–92.
- Loizeau, D., Balme, M.R., Bibring, J.P., Bridges, J.C., Fairén, A.G., Flahaut, J., Hauber, E., Lorenzoni, L., Poulakis, P., Rodionov, D., Vago, J.L., Werner, F., Westall, F., Whyte, L., Williams, R.M., 2019. Exomars 2020 surface mission: choosing a landing site. In: Lunar and Planetary Science Conference XXXIX, pp. 1–2.
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.
- McEwen, A.S., Banks, M.E., Baugh, N., Becker, K., Boyd, A., Bergstrom, J.W., Beyer, R.A., Bortolini, E., Bridges, N.T., Byrne, S., Castalia, B., Chuang, F.C., Crumpler, L.S., Daubar, I., Davatzes, A.K., Dearthoff, D.G., DeJong, A., Alan Delamere, W., Dobrea, E.N., Dundas, C.M., Eliason, E.M., Espinoza, Y., Fennema, A., Fishbaugh, K. E., Forrester, T., Geissler, P.E., Grant, J.A., Griffes, J.L., Grotzinger, J.P., Gulick, V.C., Hansen, C.J., Herkenhoff, K.E., Heyd, R., Jaeger, W.L., Jones, D., Kanefsky, B., Keszthelyi, L., King, R., Kirk, R.L., Kolb, K.J., Lasco, J., Lefort, A., Leis, R., Lewis, K. W., Martinez-Alonso, S., Mattson, S., McArthur, G., Mellon, M.T., Metz, J.M., Milazzo, M.P., Milliken, R.E., Motazedian, T., Okubo, C.H., Ortiz, A., Philippoff, A.J., Plassmann, J., Polit, A., Russell, P.S., Schaller, C., Searls, M.L., Spriggs, T., Squyres, S.W., Tarr, S., Thomas, N., Thomson, B.J., Tornabene, L.L., Van Houten, C., Verba, C., Weitz, C.M., Wray, J.J., 2010. The high resolution imaging science experiment (HiRISE) during MRO's primary science phase (PSP). *Icarus* 205, 2–37. <https://doi.org/10.1016/j.icarus.2009.04.023>.
- Palafox, L.F., Hamilton, C.W., Scheidt, S.P., Alvarez, A.M., 2017. Automated detection of geological landforms on Mars using convolutional neural networks. *Comput. Geosci.* 101, 48–56. <https://doi.org/10.1016/j.cageo.2016.12.015>.
- Papandreou, G., Chen, L., Murphy, K.P., Yuille, A.L., 2015. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: The IEEE International Conference on Computer Vision (ICCV), pp. 1742–1750.
- Parente, M., Makarewicz, H.D., Bishop, J.L., 2011. Decomposition of mineral absorption bands using nonlinear least squares curve fitting: application to Martian meteorites and CRISM data. *Planet. Space Sci.* 59, 423–442. <https://doi.org/10.1016/j.pss.2011.01.009>.
- Pina, P., Saraiva, J., Antunes, J., Bandeira, L., 2008. Automatic recognition of diverse types of polygons on Mars. In: Lunar and Planetary Science XXXIX, p. 2.
- Quantin-Nataf, C., Carter, J., Mandon, L., Balme, M., Fawdon, P., Davis, J., Thollot, P., Dehouck, E., Pan, L., Volat, M., Millot, C., Breton, S., Loizeau, D., Vago, J.L., 2019. ExoMars at Oxia Planum: probing the aqueous related Noachian environments. In: Ninth International Conference on Mars, p. 6317.
- Read, N., Woods, M., Karachalios, S., 2018. Novelty or Anomaly Hunter - driving next-generation science autonomy with large high quality dataset collection. In: I-SAIRAS, p. 2018.
- Rothrock, B., Kennedy, R., Cunningham, C., Papon, J., Heverly, M., Ono, M., 2016. SPOC: deep learning-based terrain classification for mars rover missions. In: AIAA SPACE, p. 2016. <https://doi.org/10.2514/6.2016-5539>.
- Salamunicar, G., Loncaric, S., Mazarico, E., 2012. LU60645GT and MA132843GT catalogues of lunar and Martian impact craters developed using a crater shape-based interpolation crater detection algorithm for topography data. *Planet. Space Sci.* 60, 236–247. <https://doi.org/10.1016/j.pss.2011.09.003>.
- Saranathan, A.M., Parente, M., 2021. Adversarial feature learning for improved mineral mapping of CRISM data. *Icarus* 355, 114107. <https://doi.org/10.1016/j.icarus.2020.114107>.
- Schwenzer, S.P., Woods, M., Karachalios, S., Phan, N., Joudrier, L., 2019. LabelMars: creating an extremely large martian image dataset through machine learning. In: LPSC 2019, p. 2.
- Silburt, A., Ali-dib, M., Zhu, C., Jackson, A., Valencia, D., Kissin, Y., Tamayo, D., Menou, K., 2019. Lunar crater identification via deep learning. *Icarus* 317, 27–38. <https://doi.org/10.1016/j.icarus.2018.06.022>.
- Simpson, R., De Roure, D., 2014. 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, pp. 1049–1054. <https://doi.org/10.1145/2567948.2579215>.
- Smith, E., Zuber, M.T., Frey, V., Garvin, B., Muhleman, O., Pettengill, H., Phillips, J., S., S.C., Zwally, H.J., Banerdt, W.B., Duxbury, C., Golombek, M.P., Lemoine, G., Neumann, G.A., Rowlands, D.D., Aharonson, O., Ford, P.G., Ivanov, A.B., Johnson, C.L., McGovern, J., Abshire, B., Afzal, R.S., Sun, X., 2001. Mars orbiter laser altimeter: experiment summary after the first year of global mapping of Mars. *J. Geophys. Res.* 106, 689–722.
- Stepinski, T., Vilalta, R., 2005. Digital topography models for martian surfaces. *IEEE Geosci. Remote Sens. Lett.* 2, 260–264. <https://doi.org/10.1109/LGRS.2005.848509>.
- Stepinski, T.F., Mendenhall, M.P., Bue, B.D., 2009. Machine cataloging of impact craters on Mars. *Icarus* 203, 77–87. <https://doi.org/10.1016/j.icarus.2009.04.026>.
- Tharwat, A., 2018. Classification assessment methods. *Appl. Comput. Informatics*. <https://doi.org/10.1016/j.aci.2018.08.003>.
- Urbach, E.R., Stepinski, T.F., 2009. Automatic detection of sub-km craters in high resolution planetary images. *Planet. Space Sci.* 57, 880–887. <https://doi.org/10.1016/j.pss.2009.03.009>.
- Vago, J.L., Westall, F., Coates, A.J., Jaumann, R., Korabev, O., Ciarletti, V., Mitrofanov, I., Josset, J.L., De Sanctis, M.C., Bibring, J.P., Rull, F., Goesmann, F., Steininger, H., Goetz, W., Brinckerhoff, W., Szopa, C., Raulin, F., Edwards, H.G.M., Whyte, L.G., Fairén, A.G., Bridges, J., Hauber, E., Ori, G.G., Werner, S., Loizeau, D., Kuzmin, R.O., Williams, R.M.E., Flahaut, J., Forget, F., Rodionov, D., Svedhem, H., Sefton-Nash, E., Kminek, G., Lorenzoni, L., Joudrier, L., Mikhailov, V., Zashchirinskiy, A., Alexashkin, S., Calantropio, F., Merlo, A., Poulakis, P., Witasse, O., Bayle, O., Bayón, S., Meierhenrich, U., Carter, J., García-Ruiz, J.M., Baglioni, P., Haldemann, A., Ball, A.J., Debus, A., Lindner, R., Haessig, F., Monteiro, D., Trautner, R., Volland, C., Rebeyre, P., Gouly, D., Didot, F., Durrant, S., Zekri, E., Koschny, D., Toni, A., Visentin, G., Zwick, M., Van Winnendael, M., Azkarate, M., Carreau, C., 2017. Habitability on early Mars and the search for biosignatures with the ExoMars rover. *Astrobiology* 17, 471–510. <https://doi.org/10.1089/ast.2016.1533>.
- Wallace, I., Woods, M., 2015. Master: a mobile autonomous scientist for terrestrial and extra-terrestrial research. In: 13th Symposium on Advanced Space Technologies in Robotics and Automation. ASTRA.
- Wallace, I., Schwenzer, S.P., Woods, M., Read, N., Wright, S., Waumsley, K., Joudrier, L., 2017. Labelmars.net: crowd-sourcing an extremely large high quality martian image dataset. In: Lunar and Planetary Science XLVIII 2017, pp. 1–2.
- Wang, Y., Wu, B., 2019. Active machine learning approach for crater detection from planetary imagery and digital elevation models. *IEEE Trans. Geosci. Remote Sens.* 57, 5777–5789.
- Wang, Y., Di, K., Xin, X., Wan, W., 2017. Automatic detection of Martian dark slope streaks by machine learning using HiRISE images. *ISPRS J. Photogramm. Remote Sens.* 129, 12–20. <https://doi.org/10.1016/j.isprsjprs.2017.04.014>.
- Wilhelm, T., Geis, M., Püttchneider, J., Sievernich, T., Weber, T., Wohlfarth, K., Wöhler, C., 2020. DoMars16k: a diverse dataset for weakly supervised geomorphologic analysis on Mars. *Remote Sens.* 12, 3981. <https://doi.org/10.3390/rs12233981>.
- Woods, M., Shaw, A., Barnes, D., Price, D., Long, D., Pullan, D., Le, L., 2009. Autonomous science for an ExoMars rover – like mission. *J. F. Robot.* 26, 358–390. <https://doi.org/10.1002/rob>.
- Woods, M., Shaw, A., Wallace, I., Malinowski, M., 2015. The CHAMELEON field trial: toward efficient, terrain sensitive navigation. In: 13th Symposium on Advanced Space Technologies in Robotics and Automation. ASTRA.
- Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. In: 4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.