Icarus 345 (2020) 113719

Contents lists available at ScienceDirect

Icarus

journal homepage: www.elsevier.com/locate/icarus

Machine learning approaches for classifying lunar soils

Gayantha R.L. Kodikara *, Lindsay J. McHenry

Department of Geosciences, University of Wisconsin-Milwaukee, 3209 N. Maryland Avenue, Milwaukee, WI 53211, USA

ARTICLE INFO

Keywords:
Spectroscopy
Moon
Machine learning
Classification
Feature engineering

ABSTRACT

We examine the ability of machine learning (ML) techniques to determine the physical and mineralogical properties of lunar soil using reflectance spectra. We use the Lunar Soil Characterization Consortium (LSCC) dataset to train and assess the predictive power of classification models based on their type (Mare soil and Highland soil), particle size, maturity, and the dominant type of pyroxene (High-Ca and Low-Ca). Nine ML algorithms including linear methods, non-linear methods, and rule-based methods (three from each) were selected, representing a range of characteristics such as simplicity, flexibility, computational complexity, and interpretability along with their ability to handle different types of data. Fifteen spectral parameters were initially introduced to the models as input features and a maximum of four features was selected as the best feature combinations to classify lunar soils based on their types, particle size, maturity, and the type of pyroxene. The Support Vector Machine with radial basis function (*svmRadial*) and the penalized logistic regression model (*glmnet*) performed well for all target variables with high accuracies. Band depths and Integrated band depths at 1 μm, 1.25 μm and 2 μm, band position of the 1 μm band, along with four band ratios (band tilt, band strength, band curvature, and olivine/pyroxene) are important features for classifying soil type, grain size, maturity, and type of pyroxene from reflectance spectra. This study shows that proper preprocessing and feature engineering techniques are crucial for high performance of the predictive models.

1. Introduction

The Moon is the only planetary body other than the Earth for which we have geological samples from known terrains for detailed studies. However, samples returned by the Apollo and Luna missions (1969–1976) are limited to a few small, non-random locations on the near side of the Moon. Remote Sensing provides a global perspective that the lunar samples and lunar meteorites alone cannot. To this end, a series of orbiter missions by different space agencies, each with its own tools for spectroscopy, has provided multispectral and hyperspectral data for much of the moon's surface. The US National Aeronautics and Space Administration's (NASA) Clementine mission launched in 1994 and was able to identify noritic, gabbroic/basaltic, feldspathic, and olivine-gabbro dominated areas using data from five bands in the ultraviolet-visible (UV-VIS) (e.g. Pieters et al., 2001). The European Space Agency's (ESA) first spacecraft to the Moon, Small Mission for Advanced Research and Technology (SMART-1), launched in September 2003 and carried seven instruments including a high resolution NIR point spectrometer in the wavelength range of 0.93 μm to 2.4 μm. This instrument had significantly better capabilities to discriminate between different mineralogies compared to the six filters of Clementine in the same wavelength range. In 2007, China and Japan launched two lunar exploration spacecraft named Chang' e-1 and SELENE (Selenological and Engineering Explorer). Chang'e-1 was the first step of the three-phase Chinese Lunar Exploration Program and had eight scientific instruments including an interference imaging spectrometer (IIM). Wenxiang et al. (2019) used IIM data to derive the major oxides and Mg# of the lunar surface with a resolution of ~ 200 m. Multiband Imager (MI) and Spectral Profiler (SP) aboard the SELENE (Kaguya) spacecraft was used to map the geology of the lunar surface (e.g. Hareyama et al., 2019; Trang and Lucey, 2019). The Moon Mineralogy Mapper (M^3), onboard India's first mission to the Moon, Chandrayaan-1 which launched in 2008, provided 85 bands of hyperspectral image data for the moon in the Visible- Near infrared range (e.g. Cheek et al., 2013).

The reflection behavior of a planetary surface is primarily determined by its chemical and mineralogical composition and physical properties. Electronic transition and scattering of incident light produce diagnostic absorption bands at visible to near-infrared wavelengths. Accurate characterization of the chemical and mineralogical composition of the lunar surface based on spectroscopic data from orbital

* Corresponding author.
E-mail address: gayantha_kodikara@yahoo.com (G.R.L. Kodikara).

https://doi.org/10.1016/j.icarus.2020.113719
Received 21 May 2019; Received in revised form 6 February 2020; Accepted 25 February 2020
Available online 6 March 2020
0019-1035/

platforms is critical for our understanding of the moon's origin and geological evolution. The carefully coordinated compositional and spectral investigations produced by the Lunar Soil Characterization Consortium (LSCC) database developed from the lunar highland and mare soil samples collected during the Apollo missions provide an invaluable resource to predict and validate different models developed to map the lunar surface on a global scale using orbital spectroscopic data.

Numerous empirical, theoretical, linear and non-linear statistical methods have been proposed and tested to determine the composition of the lunar soil from its optical properties (Cahill et al., 2010; Denevi et al., 2008; Li, 2006; Li and Li, 2011; Li et al., 2012; Liu et al., 2015; Noble et al., 2006; Pieters et al., 2002, 2006; Tompkins et al., 1994). However, determining physical and chemical characteristics of lunar soils from reflectance spectroscopy is not an easy task due to a complex combination of poorly understood processes and products. Some of these processes affect the behavior of reflectance spectra, including the nature of mineral mixing, the effects of particle size and degree of compaction, variation in viewing and illumination geometries, surface roughness, and space weathering introduced by solar wind sputtering and implantation and micro meteoritic bombardments (Hapke, 1993; Liu et al., 2015; Mustard and Pieters, 1989).

1.1. Previous models

Here we briefly discuss the major approaches previously applied to the field spectral and statistical modeling of lunar soil using LSCC data. Spectral modeling techniques, such as radiative transfer models (RTM) developed by Hapke (1993), use a nonlinear mixing model combined with inputs of plausible mineralogy and basic chemistry to calculate the best spectral fit for lunar soils. Denevi et al. (2008) and Cahill et al. (2010) used RTM to determine composition from modeled reflectance data, and the results were evaluated using real LSCC data. Later, Li and Li (2011) used Newton's method and the least square optimization method to solve nonlinear equations of RTM to quantify lunar surface minerals, particle sizes, and the abundance of submicroscopic metallic Fe (SMFe) created by space weathering from the LSCC data. The equation used in RTM describing the effects of submicroscopic metallic iron (SMFe) only represents the effect of smaller size SMFe. Therefore, an improved radiative transfer model was introduced by Liu et al. (2015) which considered the effects of both smaller size SMFe on the rims of soil grains and larger size SMFe in agglutinitic glass. Improved Hapke's RTM could predict the abundance of agglutinitic glass, pyroxene, and plagioclase for both immature and mature lunar soil with high accuracies. However, it was unable to predict the abundance of olivine, ilmenite, and volcanic glass with high accuracies. Spectral deconvolution techniques, such as the Modified Gaussian Method (MGM) developed by Sunshine et al. (1990) were also applied to retrieve chemical and mineralogical information from the LSCC spectra (Noble et al., 2006). This method works well for identifying both high-Ca and low-Ca pyroxene from the LSCC spectra, however it cannot predict the abundance of minerals that lack transition elements such as iron-free plagioclase or pure forsterite (Sunshine and Pieters, 1993; Noble et al., 2006; Li et al., 2012; Kodikara et al., 2016).

Pieters et al. (2006) used several statistically optimized simple linear formulations to determine the closest correlation between selected compositional parameters and an empirical combination of spectral albedo of the LSCC lunar sample spectra after resampling five Clementine bands. They concluded that no single formulation is statistically best for all parameters. Pieters et al. (2002) proposed a Principle Component Regression (PCR) approach to estimate the chemical parameters such as Fe and Ti concentrations and the I_s/FeO ratio from LSCC data (The Ferromagnetic Resonance (FMR) values for nanophase Fe0 are referred to as I_s). The adopted method was able to predict the abundance of pyroxene with high accuracy from mare soil spectra, but had low accuracy for predicting TiO_2 or ilmenite. Multiple linear regression methods have been used to map the abundance of TiO_2 and FeO, pyroxene content, degree of maturity (I_s/FeO), and the characteristic size of particles using LSCC and Clementine data (Pieters et al., 2002). Li (2006) adopted the Partial Least Squares (PLS) regression method along with PCR to predict the soil chemistry (FeO, TiO_2, Al_2O_3, SiO_2, and MgO) and mineralogy (agglutinate, pyroxene, plagioclase, olivine, ilmenite and volcanic glass) form LSCC data. He concludes that the PLS method performs better than the PCR method except for predicting volcanic glass. Hybrid partial least squares regression and back propagation neural network (PLS-BPNN) methods are also proposed and evaluated with the Genetic algorithm-partial least squares (GA-PLS) method to measure its accuracy and effectiveness by estimating the mineral and chemical abundances from LSCC data (Li et al., 2012). The results show that PLS-BPNN methods performed better than the PLS and GA-PLS methods, overcoming some limitations of PLS. In addition, PLS-BPNN was able to accommodate the spectral effects resulting from variations in particle size.

The review of the previous studies demonstrates the strengths and limits of each method for determining the mineralogical, chemical, and physical properties of the lunar soil samples using their reflectance spectra. It also shows that the combination of different methods performs better than individual methods. In some cases, linear models performed well (e.g. when the relationship between the outcome and the predictors are linear, when the decision boundaries are linear), while in other cases non-linear models performed better (e.g. when the relationship between the outcome and the predictors are non-linear, when the decision boundaries are non-linear). From these studies it can be inferred that the statistical learning methods performed better than the simple statistical methods. Therefore, in this study we test the ability of linear, non-linear and rule-based machine learning (ML) algorithms to classify select mineralogical and physical parameters of lunar soil using reflectance spectra. We also use the same LSCC dataset to train and asses the prediction power of classification models based on their type (Mare soil and Highland soil), particle size, maturity, and the dominant type of pyroxene (High-Ca and Low-Ca).

Machine learning (ML), a subset of artificial intelligence (AI), is the study of computer algorithms capable of learning to improve their performance of a task based on their own previous experience (Jordan and Mitchell, 2015). Advancements in ML over the past two decades across multiple disciplines make it an excellent technique to help find patterns in large and complex datasets. Results of various machine learning models could be functions, rules, equations, relations, probability distributions and other knowledge representations (Kononenko and Kukar, 2007).

1.2. Objectives

This study aims to (1) select the best combination of spectral parameters to identify physical and mineralogical properties of the lunar soil using LSCC data, (2) explore the best machine learning method/s to classify lunar soils according to the physical and mineralogical properties of the selected spectral parameters, and (3) to develop a reliable spectral composition model to map lunar surface using hyperspectral (e.g. Moon Mineralogy Mapper) image data.

2. Methods

The methodology includes preprocessing of spectral data, extraction of spectral features, feature engineering, model selection and computations, evaluation, and selection of the best models. The entire study was conducted using the free and open source R statistical package (www.r-project.org) along with Rstudio (https://www.rstudio.com).

2.1. Dataset

The Lunar Soil Characterization Consortium (LSCC) data set includes

nine mare and ten highlands soil samples collected by Apollo astronauts and analyzed in the laboratory using reflectance spectroscopy to study space weathering effects on lunar soils (Taylor et al., 2001, 2010). The dataset represents a broad suite of soils representing both the compositional diversity and degree of maturity found in lunar soils. Within this dataset, each soil sample was subdivided into four particle size groups (<10 μm, 10–20 μm, 20–45 μm, < 45 μm). The bidirectional reflectance spectra of the <10 μm, 10–20 μm, and 20–45 μm particle size groups were measured over the spectral range 0.3–2.6 μm with 5 nm sampling resolution with an incident angle of 30^0, emittance angle 0^0, and phase angle 30^0. Mineral abundances, maturity, and chemistry were also determined using other methods (Pieters et al., 2006; Taylor et al., 2001, 2010). All spectral data and compositions measured under LSCC are accessible at http://www.planetary.brown.edu/relabdocs/LSCCsoil.html.

Selected attributes for this study include the type of the lunar sample (Highland or Mare), soil size fraction, measurements of I_s/FeO (a measure of maturity), and modal percentage of four types of pyroxene (Orthopyroxene, Pigeonite, Mg-clinopyroxene and Fe-pyroxene) (Pieters et al., 2006; Taylor et al., 2001, 2010). The I_s/FeO is the amount of single-domain nanophase Fe^0 in the soil normalized to the FeO content, where I_s is measured using the ferromagnetic resonance (FMR) technique in the Magnetics Lab of R. V. Morris, at Johnson Space Center (Morris, 1976, 1978; Taylor et al., 2001). We have modified the original sample names, adding an "H" or "M" at the beginning to identify each as a Highland or Mare sample, and appending an A, B, or C at the end to distinguish grain size (A = <10 μm soil fraction, B = size fraction between 10 and 20 μm, and C = size fraction 20–45 μm). Table 1 shows the

Table 1
Basic attributes of the lunar soils used for this analysis, from the LSCC database.

Sample	Type	Size (μm)	I_s/FeO	Ortho (%)	Pigeo (%)	Mg_Clino (%)	Fe_Pyro (%)
H14141A	Highland	< 10	14.5	3.37	4.29	2.19	0.44
H14141B	Highland	10–20	11.6	4.07	4.58	1.85	0.38
H14141C	Highland	20–45	5.8	7.57	8.08	3.08	1.08
H14163A	Highland	< 10	87	1.51	1.38	0.91	0.14
H14163B	Highland	10–20	64.8	5.68	4.94	2.41	0.78
H14163C	Highland	20–45	43.2	6.5	5.66	3.1	0.92
H14259A	Highland	< 10	174.8	1.92	1.96	1.77	0.29
H14259B	Highland	10–20	101.8	3.72	3.18	1.66	0.5
H14259C	Highland	20–45	77.2	7.4	6.14	3.04	1.57
H14260A	Highland	< 10	144.9	2.58	3.15	1.51	0.48
H14260B	Highland	10–20	98.9	5.14	4.23	2	0.64
H14260C	Highland	20–45	80.2	4.68	4.99	3.07	0.94
H61141A	Highland	< 10	119.3	0.22	0.23	0.19	0.06
H61141B	Highland	10–20	81.6	1.69	2.15	1.45	0.04
H61141C	Highland	20–45	75.5	1.68	1.38	1.11	0.18
H61221A	Highland	< 10	19.8	0.56	0.55	0.37	0.02
H61221B	Highland	10–20	13.89	1.82	1.43	1.95	0.14
H61221C	Highland	20–45	8.4	2.96	2.24	1.98	0.19
H62231A	Highland	< 10	169	0.28	0.27	0.3	0.03
H62231B	Highland	10–20	109.9	1.99	1.55	1.74	0.12
H62231C	Highland	20–45	80.7	2.08	1.33	1.52	0.19
H64801A	Highland	< 10	115.2	1.18	0.84	0.64	0
H64801B	Highland	10–20	84.9	1.24	0.96	0.6	0.01
H64801C	Highland	20–45	83.4	2.03	1.15	1.33	0.01
H67461A	Highland	< 10	35.2	1.09	0.64	1.06	0.04
H67461B	Highland	10–20	23.9	1.47	1.07	1.52	0.05
H67461C	Highland	20–45	22.3	2.96	1.61	2.53	0.18
H67481A	Highland	< 10	38.5	1.38	1.05	1.41	0.05
H67481B	Highland	10–20	33	2.55	1.27	1.73	0.13
H67481C	Highland	20–45	20.7	2.95	1.54	1.94	0.17
M10084A	Mare	< 10	145	0.42	1.81	5.96	0.25
M10084B	Mare	10–20	87	0.61	3.23	7.81	0.57
M10084C	Mare	20–45	67	0.34	3.94	10.35	1.38
M12001A	Mare	< 10	115	2.07	6.06	4.79	0.58
M12001B	Mare	10–20	67	2.24	7.36	7.64	0.7
M12001C	Mare	20–45	51	0.82	8.77	7.52	1.77
M12030A	Mare	< 10	32	2.2	7.07	5.18	0.89
M12030B	Mare	10–20	17	2.86	10.18	6.59	1.75
M12030C	Mare	20–45	12	3.85	15.2	12.14	2.59
M15041A	Mare	< 10	161	0.79	2.49	1.62	0.42
M15041B	Mare	10–20	92	2.35	8.14	5.12	1.37
M15041C	Mare	20–45	66	3.77	10.48	6.75	1.49
M15071A	Mare	< 10	159	2.11	5.2	3.21	0.4
M15071B	Mare	10–20	80	2.13	7.64	5.56	1.38
M15071C	Mare	20–45	49	3.22	10.27	6.98	1.6
M70181A	Mare	< 10	104	0.59	1.76	1.98	0.31
M70181B	Mare	10–20	63	1.2	2.57	3.74	0.97
M70181C	Mare	20–45	53	1.51	4.7	8.15	1.37
M71061A	Mare	< 10	28	0.95	2.72	4.11	0.49
M71061B	Mare	10–20	14	1.32	4.07	5.97	1.12
M71061C	Mare	20–45	9	1.15	6.87	10.29	2.18
M71501A	Mare	< 10	88	1	2.85	4.1	0.76
M71501B	Mare	10–20	0.19	1.47	4.61	6.34	1.25
M71501C	Mare	20–45	28	1.44	6.31	11.1	2.35
M79221A	Mare	< 10	169	0.6	1.42	1	0.6
M79221B	Mare	10–20	78	1.64	2.86	3.24	1.82
M79221C	Mare	20–45	57	1.47	3.72	4.85	3.14

modified name and selected attributes used for this analysis.

Highland soils start with the letter H and Mare soils start with the letter M. The number in the sample name represents the LSCC soil sample name. The last letter of the sample name represents the size of the soil fraction (A: <10 μm, B: 10–20 μm, and C: 20–45 μm). I_s/FeO is linked to maturity (see text). Ortho = Orthoclase, Pigeo = Pigeonite, Mg_Clino = Mg rich Clinopyroxene, Fe_Pyro = iron rich Pyroxene.

2.2. Pre-processing of spectral data

Downloaded LSCC spectral data (in .txt format) were converted to a comma separated values (.csv) file format and imported into the R statistical software to create a spectral database using the Hyperspectral Data Analysis (hsdar) package developed by Lehnert et al. (2018). The hsdar package consists of commonly used hyperspectral data processing functions with the existing functionality of R for a wide range of statistical analysis. The hsdar package was originally designed to calculate vegetation indices, red edge parameters, and simulation of reflectance and transmittance using the leaf reflectance model PROSPECT and the canopy reflectance model PROSAIL (Lehnert et al., 2014, 2015, 2018), however we adapted this package to create our spectral library and to conduct basic spectral processing including spectral resampling, continuum removal, and to plot the spectra. The basic attributes listed in Table 1 were also added to the spectral database as an associated attribute of each spectrum. We created a Spectral Resampling Bandpass Filter using Gaussian spectral response functions defined by the fwhm (full-width-half-maximum) values of the Moon Mineralogy Mapper (M^3) camera channels to resample our LSCC spectra to M^3 (Lundeen et al., 2010). This will later allow us to determine and apply the most suitable machine learning algorithm to map the lunar surface using M^3 hyperspectral image data.

2.3. Extraction of spectral features

The position, depth, width, area and the shape of absorption bands of a given spectrum are controlled by the composition and crystal structure of the absorbing species (Burns, 1970; Clark, 1999; Cloutis et al., 1986). Different combinations of techniques such as curve matching (e.g., Clark, 1999), empirical curve fitting (e.g., Denevi et al., 2008) and curve deconvolution (e.g., Sunshine et al., 1990) are used to derive this physical, chemical, and mineralogical information from the spectra. Curve matching is mostly accomplished by visually comparing one spectrum to another. Simple band rationing methods are also used to match spectra based on their spectral characteristics (Borst et al., 2012; Dhingra, 2008; Mouelic and Langevin, 2001; Sivakumar et al., 2017). These methods may ignore other factors that can change the spectrum of a particular mineral, or an assemblage of mineral spectra. In empirical curve fitting, spectra are systematically curve fit using a polynomial fit about the minima or centers of absorption to derive these metrics. Then these minima or centers are compared to similar curve fits performed using laboratory spectra of calibrated standards or calibrated mixtures (McCraig et al., 2017). Empirical studies are useful as diagnostic tools for mixtures of specific minerals, but can only be applied to the minerals studied. Curve deconvolution (typically referred to as Gaussian fit optimization) is a quantitative approach to extract the absorption parameter by fitting the gaussian curves with selected spectra (Sunshine et al., 1990). Use of this method for an unknown spectrum is only dependent on the spectrum itself. Therefore, it does not require known spectra or calibrated standards or mixtures. In this work, we have used the simple band rationing methods and gaussian curve fitting methods to extract the absorption band parameters from the LSCC spectra.

We have calculated four simple band indices using the resampled spectra. Band strength (also called key ratio) and band curvature were originally developed by Tompkins and Pieters (1999) to distinguish different mafic-bearing lithologies on the lunar surface using Clementine multispectral data. The band strength, the ratio of the 1000 nm Clementine band to the Clementine band at 750 nm, is a proxy for the abundance of mafic minerals. Lunar materials with a weak 1000 nm absorption band, such as in anorthosite and mature soils, show higher band strength than the mafic minerals and un-weathered soils. Spectral curvature is defined by the angle formed in the Clementine bands at 750 nm, 900 nm and 950 nm. It is generally sensitive to the type of mafic silicate present and especially distinguishes between longer and shorter wavelength absorptions due to low- and high- Ca pyroxene spectra (Tompkins and Pieters, 1999). Lunar materials with abundant low-Ca pyroxene (noritic compositions) would have higher curvature values than high-Ca pyroxenes. Olivine rich materials would have much lower band curvature values based on this calculation (Pieters et al., 2001). Mouelic and Langevin (2001) used the intensity ratio between the 2000 nm and 1250 nm bands to detect olivine-rich areas using Clementine NIR data. This index is named “olivine/pyroxene”, since it is high for olivine-dominated rocks and low for orthopyroxene-dominated rocks. Borst et al. (2012) used these parameters to map surface mineralogy and stratigraphy of the South Pole-Aitken (SPA) basin using Clementine UV/VIS and NIR data. Recently, Sivakumar et al. (2017) used the same parameters to map the mineralogy of the lunar surface using hyperspectral M^3 data with slight modifications of the band positions.

Continuum removed spectra were used to calculate the other spectral parameters using LSCC reflectance spectra. The continuum is the background absorption onto which other absorption bands are superimposed (Clark et al., 1990). The continuum removal transformation is performed by establishing a continuum line connecting the maximum points of the reference spectrum. Two kinds of continuum removal methods are commonly used; the “upper convex hull” introduced by Green and Craig (1985) and the “segmented upper hull” introduced by Clark et al. (1987). Upper convex hull is defined as the lowest convex curve lying above the given spectrum (Green and Craig, 1985). Segmented upper hull is performed on segments of a spectrum defined by local minima and maxima (Clark et al., 1987). Therefore, the main difference between these two is that the resulting continuum line in convex hull must be convex while the resulting continuum line of segmented hull can be convex or concave. Theoretically, segmented hull can identify more small absorption bands than the convex hull. We applied both methods to select the best method for our analysis. Overall, the convex hull method performed better than the segment hull method, as shown for sample M12030 in Fig. 1. The segmented hull method takes more small segments to define the continuum line, resulting in higher uncertainties (e.g. Fig. 1 (e)).

The convex hull method identifies three absorption bands for most of the LSCC type A (< 10 μm size fraction) spectra. Lunar spectra tend to be dominated by a very prominent pyroxene absorption band centered near 1 and 2 μm, and a weaker absorption band near 1.2 μm (Adams, 1974). The wavelengths at band depth of the 1 μm and 2 μm bands are highly dependent on the type of pyroxene and its composition. If orthopyroxene is the dominant component, the band center ranges from 0.90 to 0.93 μm. If the clinopyroxene is the dominant component, the band center varies from 0.98 to 1.0 μm. A broad absorption band center near 0.95 μm indicates significant amounts of both pyroxenes (Adams, 1974; Burns, 1970; Pieters and Englert, 1993). Iron content also influences the locations of the absorption bands. With an increase in the iron content, the 1 and 2 μm absorption band centers move towards longer wavelengths, along with a strengthening of the 1.2 μm absorption band (Adams, 1974). Olivine exhibits a characteristic broad three-component absorption band centered beyond 1.05 μm (Burns, 1970). Plagioclase feldspar also has a diagnostic absorption band in the 1.25 μm region (Adams, 1975; McCord et al., 1981). This absorption band is caused by electronic transitions in Fe^{2+} cations substituting for Ca^{2+} in an irregular eightfold to 12-fold site (Adams and Goullaud, 1978; Conel and Nash, 1970). Laboratory and modeling studies of mineral mixtures have shown that the plagioclase absorption band is distinguishable only if the sample has more than ~ 85 vol% of plagioclase (Crown and Pieters, 1987; Nash and Conel, 1974). Cheek et al. (2013) identified and mapped the

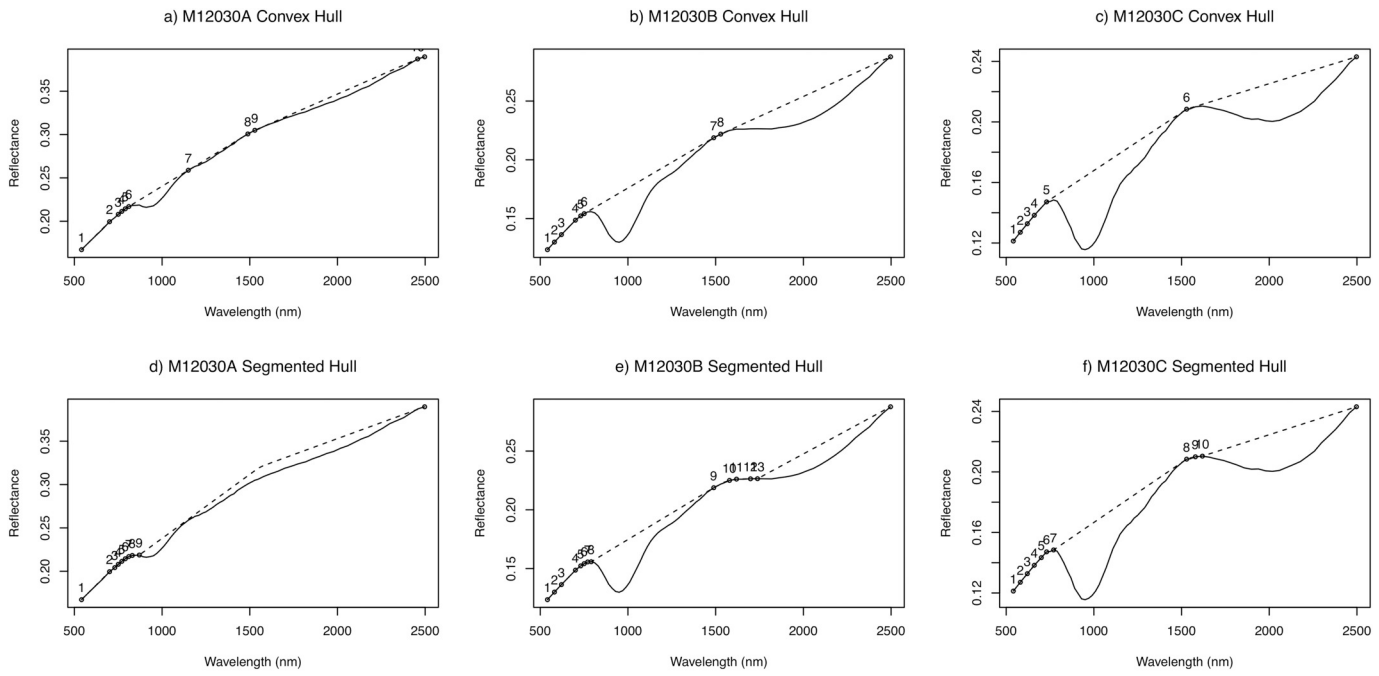


Fig. 1. Reflectance spectrum (solid line) of different soil size fractions of soil sample M12030. Continuum line calculated from Convex Hull and Segmented Hull methods is shown as a dashed line on each spectrum.

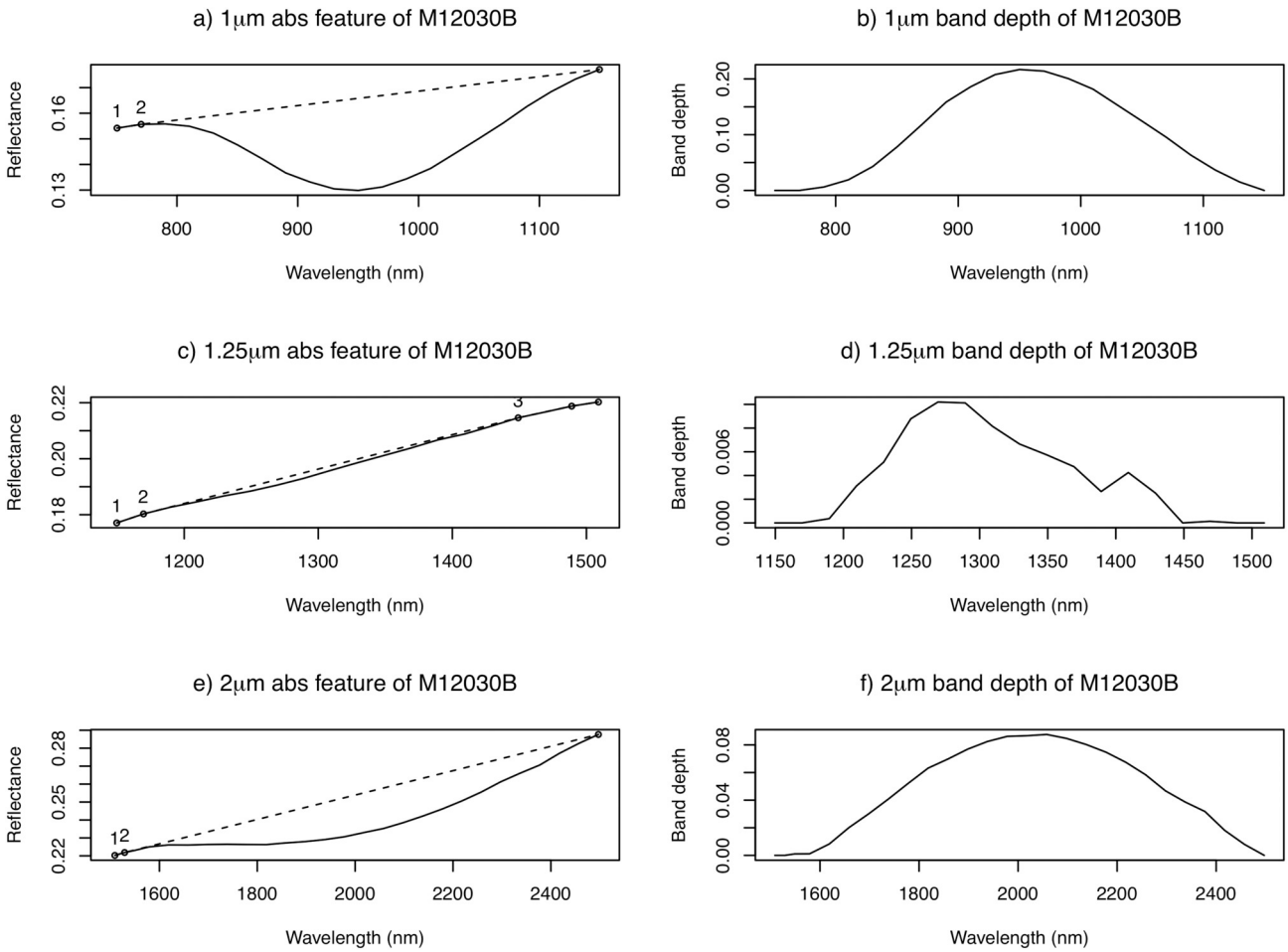


Fig. 2. Continuum line and resulting band depth for three wavelength regions of M12030B spectra.

plagioclase-dominant lithology at Orientale basin using the 1.25 μm absorption band from Moon Mineralogy Mapper data. Noble et al. (2006) also found an absorption band around 1.28 μm in the LSCC spectra (in addition to the 1 μm and 2 μm) using the Modified Gaussian Method. However, they found that plagioclase and agglutinate were not a possible cause of the 1.25 μm absorption band, and that pyroxene appears to be largely responsible for it. To address this issue, we also decided to derive the absorption band parameters from the absorption band at 1.25 μm, in addition to the absorption bands at 1 μm, 2 μm. However, neither the convex hull or the segmented hull methods were able to correctly identify the absorption band around 1.25 μm from the large size fraction of LSCC spectra. While Fig. 1 (a) correctly identified the absorption band at 1.25 μm, when the 1 μm absorption band gets stronger as the size fraction increases (Fig. 1 (b) and (c)), the two overlap. This can also happen due to the overlap of the pyroxene absorption band with the olivine absorption band. For these reasons, and after thorough spectral analysis, we applied the convex hull method for three segmented spectra with the wavelength range from 750 nm to 1150 nm, 1150 nm to 1510 nm, and 1510 nm to 2500 nm. Fig. 2 shows the three spectral regions with their spectral absorption bands for sample M12030B.

Fig. 3 shows all absorption bands extracted using this method. Note that some of the samples show strong absorption bands, while others don't show any. In addition, the 1 μm absorption bands show the highest band depths while the 1.25 μm absorption bands show the lowest, as would be expected.

Three main absorption parameters, including maximum absorption depth (referred to from here onward as band depth), wavelength at maximum absorption depth (referred to from here onward as band position) and integrated band depth (referred to from here onward as IBD) were calculated for each wavelength range, for a total of nine absorption parameters. The band depth (D) is usually defined relative to the continuum, R_c:

$$D = 1 - \frac{R_b}{R_c}$$

where R_b is the reflectance at the band bottom and R_c is the reflectance of the continuum line at the same wavelength as R_b (Clark and Roush, 1984). The absorption depth is related to the abundance of the absorber and the grain size of the material concerned. The band position is the wavelength of R_b. It changes with the composition of the material concerned. Several recent studies along with the initial studies by the M^3 team have shown that the IBD of the crystal field absorptions at 1 and 2 μm provide an excellent summary of the mineral diversity of the lunar surface using M^3 data (Cheek et al., 2013; Cheek et al., 2011; Mustard et al., 2011). IBD is defined as the sum of band depths in the selected wavelength region relative to the local continuum R_c. As an example, IBD at the 1 μm absorption band ($IBD1\mu m$) is defined by:

$$IBD1\mu m = \sum_{n=7}^{27} -\frac{R_b(750+20n)}{R_c(750+20n)}$$

In this equation, 750 is the first wavelength (nm) in a series for integration, 20 specifies the wavelength interval in nanometers, and n is the number of channels over which the integration is performed (Cheek et al., 2013).

Cloutis et al. (1986) showed that the ratio of the absorption band

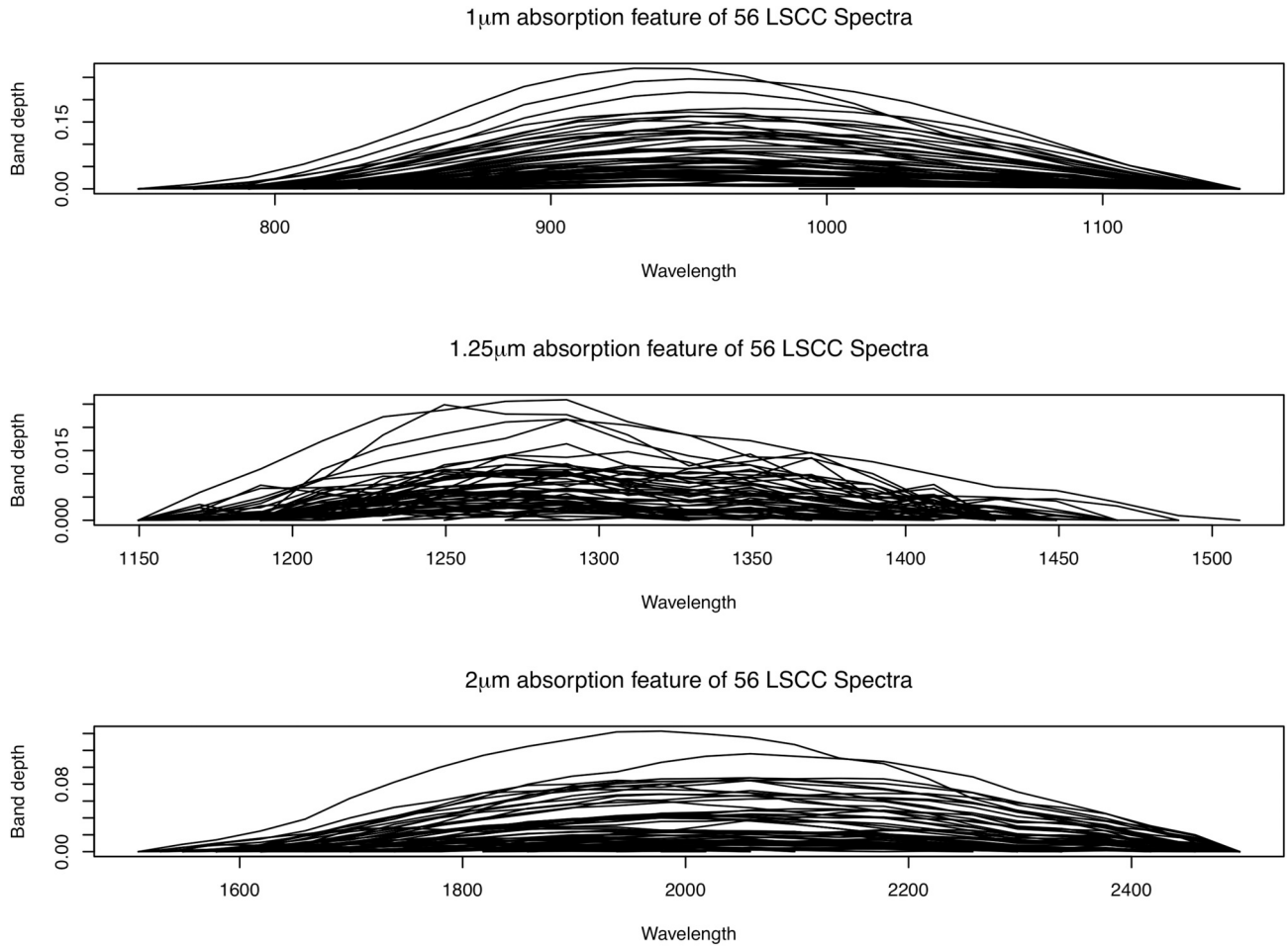


Fig. 3. Band depths of spectra in three wavelength regions.

area at 1 μm and 2 μm is a sensitive indicator of olivine- orthopyroxene abundance and is almost independent of particle size and mineral composition. The Rationing technique enhances the expression of spectral absorption bands by suppressing artifacts that result from residual and systematic instrument errors common to all spectra in a data set (Cloutis et al., 1986; Mustard et al., 2011). Therefore, the ratio of band depth at 1 μm and 2 μm, and the ratio of IBD at 1 μm and 2 μm was also calculated. A total of 15 spectral indices from the LSCC spectra were derived to use as the input data for machine learning models (Table 2).

Taylor et al. (2010) classified lunar soil maturity into three classes based on I_s/FeO values of each sample: Immature (I_s/FeO < 30), Submature (30 < I_s/FeO > 60), and Mature (I_s/FeO > 60). The drawback of this fixed-width binning is that it can produce irregular bins that are not uniform based on the number of observations that fall into each bin, which can negatively affect the performance of the model. Adaptive binning is a safer strategy in this case, in which we let the data decide the optimal number of bins and optimal boundaries for each bin (Kononenko and Kukar, 2007; Kuhn and Johnson, 2013). We used the Ward's hierarchical clustering method to classify the dataset based on the I_s/FeO values. Hierarchical cluster analysis is a statistical method for finding relatively homogeneous clusters of observations based on dissimilarities or distances between objects (Ward, 1963). It starts with each observation as a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. We have assigned two classes (Mature and Immature) based on the dendrogram and clustering results (Table 3). We also combined the orthopyroxene and pigeonite abundances into a "low-Ca pyroxene" category and the Mg-clinopyroxene and Fe-pyroxene into a "high-Ca pyroxene" category. In this case, we neglected the effect of iron in determining the band centers because the type of pyroxene is more important for determining the band centers than is iron content in lunar soils (Noble et al., 2006). The relative abundance of low-Ca and high-Ca pyroxene was used to categorize samples into two pyroxene classes named Low (low-Ca pyroxene prominent) and High (high-Ca pyroxene prominent) (Table 3).

2.4. Feature engineering

In previous studies, the input variables created to make a prediction are variously called features, predictors, independent variables, regressors, attributes, or sometimes just variables (Bowles, 2015; James et al., 2017; Kuhn and Johnson, 2013). The variable being predicted is also referred to by different names, such as target, label, response, outcome or dependent variables (Bowles, 2015; James et al., 2017; Kuhn and Johnson, 2013). Throughout this text, we will use the term "features" for input variables and "target" for the variable being predicted. Therefore, we have fifteen features listed in Table 2 and four target variables called Type (soil type; Mare or Highland), Size (Soil size fraction), Maturity (Maturity of the soil: Mature or Immature) and Pyroxene (the type of pyroxene: High-Ca or Low-Ca). The process of transforming and combining the original input data into features that better represent the underlying problem to the selected predictive models is called feature engineering (Bowles, 2015; Kuhn and Johnson, 2013). This process includes pre-processing of existing features, adding new features, and selecting the best features or combination of features based on feature importance.

Pre-processing steps are determined by the type of models being used and commonly used steps include: handling missing values, centering and scaling if the features are skewed, removing noise or outliers, removing features with the highest collinearity, and binning features based on the characteristics of the data (Kononenko and Kukar, 2007; Kuhn and Johnson, 2013). In many cases, some features have missing values. Missing values can either be ignored, replaced by the most probable value, or treated using the probability distribution of other values (Kononenko and Kukar, 2007). Fortunately, missing values were not observed in our dataset. Spectral parameters derived from M10084A were removed from the database, since it is an outlier under the multivariant Cook's distance approach (Cook, 1977). Cook's distance is calculated by removing the i^{th} data point from the model and recalculating the regression. It summarizes how much all of the values in the regression model change when the i^{th} observation is removed. The Cook's distance (D_i) is defined by:

$$D_i = \sum_{j=1}^{n} \frac{\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE}$$

where, $\hat{Y}_j$ is the value of j^{th} fitted response when all the observations are included, $\hat{Y}_{j(i)}$ is the value of j^{th} fitted response, where the fit does not include observation *i*, *p* is the number of fitted parameters, and *MSE* is the root mean square error.

Skewness of the dataset also decreases the predictive performance of the models. Skewness (S_k) is the degree of distortion from the symmetrical normal distribution in a set of data.

$$S_k = \frac{\sum (x_i - \bar{x})^3}{(n-1)v^{3/2}}$$

where,

$$v = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

where *x* is the feature variable, n is the number of features, and $\bar{x}$ is the sample mean of the features (Kuhn and Johnson, 2013). Skewness values of all features were calculated, and the results show acceptable skewness values for all except for highly positively skewed Ratio of Integrated Band Depth (RIBD). Replacing the data with log, square root or

Table 2
Spectral parameters used in this analysis with their response to mineralogy based on the literature.

Name	Parameter	Olivine	Low-Ca pyroxene	High-Ca pyroxene	Anorthosite
Max1um	Maximum absorption depth 1 μm absorption band		+	+	
MaxWl1um	Wavelength at maximum depth of 1 μm absorption band		−	+	
IBD1um	Integrated Band Depth at 1 μm absorption band	+	+	+	
Max1.25um	Maximum absorption depth 1.25 μm absorption band				+
MaxWl1.25um	Wavelength at maximum depth of 1.25 μm absorption band				+
IBD1.25um	Integrated Band Depth at 1.25 μm absorption band				+
Max2um	Maximum absorption depth 2 μm absorption band		+	+	
MaxWl2um	Wavelength at maximum depth of 2 μm absorption band		−	+	
IBD2um	Integrated Band Depth at 2 μm absorption band		+	+	
BCU	Band Curvature [(R_{750}/R_{910}) + (R_{1008}/R_{910})]	+	+		+
BST	Band Strength [R_{1008}/R_{750}]		+	+	+
BTL	Band Tilt [R_{910}/R_{1008}]		+	+	
OVP	Olivine/ Pyroxene [R_{2018}/R_{1008}]	+			
RBD	Ratio of band depth at 1 μm and 2 μm		+	+	
RIBD	Ratio of Integrated band depth at 1 μm and 2 μm				

Table 3
Basic characteristics of the lunar soils derived from Table 1.

Sample	Is/FeO	Maturity	Low_Ca (%)	High_Ca (%)	Pyroxene
H14141A	14.50	Immature	7.66	2.63	Low
H14141B	11.60	Immature	8.65	2.23	Low
H14141C	5.80	Immature	15.65	4.16	Low
H14163A	87.00	Mature	2.89	1.05	Low
H14163B	64.80	Immature	10.62	3.19	Low
H14163C	43.20	Immature	12.16	4.02	Low
H14259A	174.80	Mature	3.88	2.06	Low
H14259B	101.80	Mature	6.90	2.16	Low
H14259C	77.20	Mature	13.54	4.61	Low
H14260A	144.90	Mature	5.73	1.99	Low
H14260B	98.90	Mature	9.37	2.64	Low
H14260C	80.20	Mature	9.67	4.01	Low
H61141A	119.30	Mature	0.45	0.25	Low
H61141B	81.60	Mature	3.84	1.49	Low
H61141C	75.50	Mature	3.06	1.29	Low
H61221A	19.80	Immature	1.11	0.39	Low
H61221B	13.89	Immature	3.25	2.09	Low
H61221C	8.40	Immature	5.20	2.17	Low
H62231A	169.00	Mature	0.55	0.33	Low
H62231B	109.90	Mature	3.54	1.86	Low
H62231C	80.70	Mature	3.41	1.71	Low
H64801A	115.20	Mature	2.02	0.64	Low
H64801B	84.90	Mature	2.20	0.61	Low
H64801C	83.40	Mature	3.18	1.34	Low
H67461A	35.20	Immature	1.73	1.10	Low
H67461B	23.90	Immature	2.54	1.57	Low
H67461C	22.30	Immature	4.57	2.71	Low
H67481A	38.50	Immature	2.43	1.46	Low
H67481B	33.00	Immature	3.82	1.86	Low
H67481C	20.70	Immature	4.49	2.11	Low
M10084A	145.00	Mature	2.23	6.21	High
M10084B	87.00	Mature	3.84	8.38	High
M10084C	67.00	Immature	4.28	11.73	High
M12001A	115.00	Mature	8.13	5.37	Low
M12001B	67.00	Immature	9.60	8.34	Low
M12001C	51.00	Immature	9.59	9.29	Low
M12030A	32.00	Immature	9.27	6.07	Low
M12030B	17.00	Immature	13.04	8.34	Low
M12030C	12.00	Immature	19.05	14.73	Low
M15041A	161.00	Mature	3.28	2.04	Low
M15041B	92.00	Mature	10.49	6.49	Low
M15041C	66.00	Immature	14.25	8.24	Low
M15071A	159.00	Mature	7.31	3.61	Low
M15071B	80.00	Mature	9.77	6.94	Low
M15071C	49.00	Immature	13.49	8.58	Low
M70181A	104.00	Mature	2.35	2.29	Low
M70181B	63.00	Immature	3.77	4.71	High
M70181C	53.00	Immature	6.21	9.52	High
M71061A	28.00	Immature	3.67	4.60	High
M71061B	14.00	Immature	5.39	7.09	High
M71061C	9.00	Immature	8.02	12.47	High
M71501A	88.00	Mature	3.85	4.86	High
M71501B	0.19	Immature	6.08	7.59	High
M71501C	28.00	Immature	7.75	13.45	High
M79221A	169.00	Mature	2.02	1.60	Low
M79221B	78.00	Mature	4.50	5.06	High
M79221C	57.00	Immature	5.19	7.99	High

inverse values may help to remove the skewness (Kuhn and Johnson, 2013). In this case, the square root of RIBD (named here as RIBD2) gave the better result than RIBD (Fig. 4). Therefore, the predictor RIBD was replaced by RIBD2.

The main goal of using the Machine Learning (ML) approach in this study is to select the best computational model to predict some properties of lunar soils using reflectance spectra. Therefore, we first need to train the selected computational models using known data. At the end of the training process, the correct model should predict correct outputs for the input training data, and it should also be able to generalize well to previously unseen data (Kononenko and Kukar, 2007; Kuhn and Johnson, 2013; Reitermanova, 2010). If the model generalized poorly (overtrained), it just memorizes the training data and it will not be able to give correct output for the validation data (data which were not used to train the model). Therefore, good prediction and good generalization (also called minimum bias and minimum variance) are the two crucial measures of a good ML method (Kononenko and Kukar, 2007). Cross Validation (CV) techniques are the most common techniques used to balance between minimal bias and minimal variations in the models. Cross-validation techniques can also be used to evaluate and compare more models, various training algorithms, or to find optimal model parameters (Kononenko and Kukar, 2007). The two most commonly used CV methods are hold-out cross validation and k-fold cross validation. Hold-out cross-validation separates the dataset into three mutually disjointed subsets, called training, validation, and testing. The model is trained on the training subset, while the validation subset is periodically used to evaluate the model performance during the training to avoid over-training. When the performance on the validation dataset is good enough, the testing subset is used to validate the models' performance (Reitermanova, 2010). The k-fold cross-validation (k-fold CV) uses a combination of more tests to obtain a stable estimate of the model performance (Kuhn and Johnson, 2013).

An improper split of the dataset can lead to an excessively high variance in the model performance. Some data splitting methods are simple and widely used, such as simple random sampling (SRS). SRS does not control for any of the features, such as the percentage of observations in each class. Therefore, it may suffer from high variance of the model performance. Other methods are deterministic, such as systematic sampling. Systematic sampling is designed for naturally ordered datasets and is restricted to specific types of datasets (e.g. time series). The more sophisticated methods, such as stratified sampling, exploit the structure of the data to reach confident results at the expense of higher computational costs (Cochran, 1977). The stratified random sampling method, the method used in this study, selects samples from each cluster with a uniform probability. There are three basic ways to choose the number of samples to be selected per cluster: equal allocation, proportional allocation, and optimal allocation (Reitermanova, 2010). Equal allocation takes the same number of samples from each cluster, while the proportional allocation selects samples based on the size of each cluster. In optimal allocation, the number of samples to be selected for each cluster will depend not only on the cluster size, but also on the standard deviation of the samples in the cluster (Cochran, 1977). We used the stratified random sampling with proportional allocations to split the entire dataset into two sets as training (to train the models) and validation (to evaluate their performance). Training sets include 80% of the observations (45 observations) and the rest are assigned as validation (11 observations).

Correlations between all features were also measured to see which data have highly correlated features. The redundant features frequently add more complexity to the model than information. They can also result in highly unstable models, numerical errors, and poor performance in some predictor models, such as in linear regression. Fig. 5 shows the correlation matrix of the calculated features. Each pairwise correlation is computed from the training dataset and colored based on their magnitudes. Dark blue color indicates strong positive correlations, dark red is used for strong negative correlation, and white indicates no empirical relationship between the features.

The feature variables in the figure are grouped using a clustering technique. Therefore, the figure shows one big block of strong positive correlations as a cluster of collinearity. It shows the correlation between band depth and IBD of the 1 μm band with the same indices of the 2 μm band. The Band depth of the 1.25 μm band is highly correlated with its IBD. These correlations are expected, since band depth and IBD are two properties of the same absorption band which in this case follows a gaussian distribution. RBD and RIBD2 are also relatively highly correlated, since they represent the ratio of band depth and their derivatives. Additionally, BCU, Max1um, IBD1um, Max2um, and IBD2um are highly negatively correlated with BST. Band Strength (BST) is approximately opposite of band depth for the 1 μm absorption band. Since the 1 μm and 2 μm absorption bands are highly correlated with each other, BST should

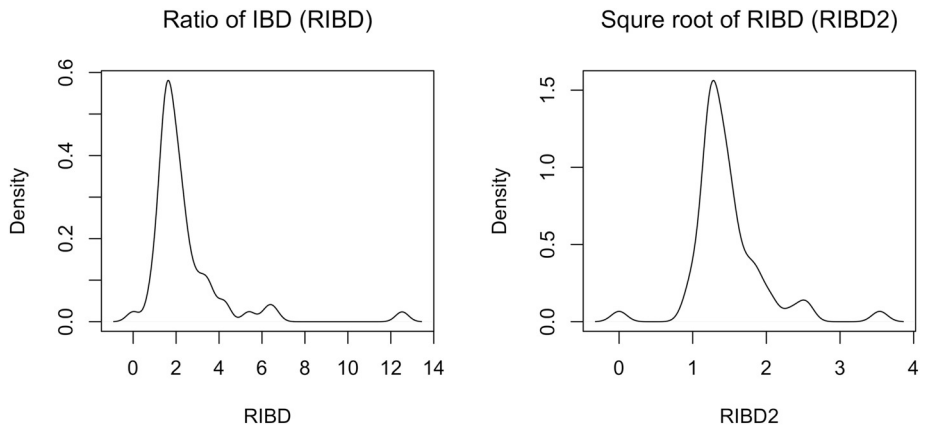


Fig. 4. Distributions of feature attributes of Ratio of Integrated Band Depth (RIBD) and its square root (RIBD2).

also be highly negatively correlated with the 2 μm absorption band. Band curvature is also related to the band depth of the absorption band based on their algorithms. Therefore, it is also negatively correlated with BST. At this stage, we decided to keep all features, because the method applied here identified only the collinearities in two dimensions and these features could have significant positive effects on the performance of some high dimensional predictive models.

Feature selection is an effective way to identify the features in the dataset that are important and discard others as irrelevant and redundant. Feature selection techniques can be divided into three different categories based on the relation between the selection scheme and the basic induction algorithm. Those categories include filter, wrapper, and embedded (Blum and Langley, 1997). For filtering methods, the preprocessing steps use the general characteristics of the training set to select some features and to exclude others. The wrapper approach uses a learning algorithm as a "black box" along with a resampling technique to select the best features according to some predictive measures. In the embedded approach, the features are selected in the process of learning (Blum and Langley, 1997). Here we use the Learning Vector Quantization (LVQ) model to estimate the feature importance. LVQ is a family of algorithms for statistical pattern classification, which aims to learn codebook vectors representing class regions. Even if the class distributions of the input samples would overlap at the class borders, the codebook vectors of each class will stay within each class of region at all times (Kohonen, 2001). Feature importance scores and the first six features selected for each class are shown in Fig. 6.

Fig. 6 (a) shows the importance of various features for soil type classification. Band position at the 1 μm absorption band and the olivine/pyroxene ratio (OVP) provide the highest ability to distinguish highland soils from mare soils. Conditional density plots of the soil type category in Fig. 6 (b) show overlapping of feature classes except for the feature plots of MaxWl1um and OVP. Several other studies also identified OVP as a good parameter to distinguish Highland and Mare samples based on the presence of olivine or mafic intrusive rocks (Borst et al., 2012; Sivakumar et al., 2017). Band depth at 1.25 μm and its IBD are the fourth and fifth most important features, probably due to the presence of olivine in Mare soil samples (Sunshine et al., 1990). If this were due to anorthite, highland soils should have higher band depth and IBD than the mare soils. Noble et al. (2006) reached the same conclusion based on the MGM analysis. Band strength was the least important feature for classifying lunar soils based on their type.

Fig. 6 (c) ranks feature importance for classifying lunar soils based on the grain size fraction. The IBD of the 1 μm and 2 μm bands, and the band depth of the 2 μm band are the most important for classifying lunar soils based on their size fractions. Feature plots of IBD2μm and Max2μm (Fig. 6 (d)) also show very similar patterns, as discussed for the correlation plot (Fig. 5). All of the feature plots except BST in Fig. 6 (d) show similar patterns since all represent the band depth at 1 μm and 2 μm,

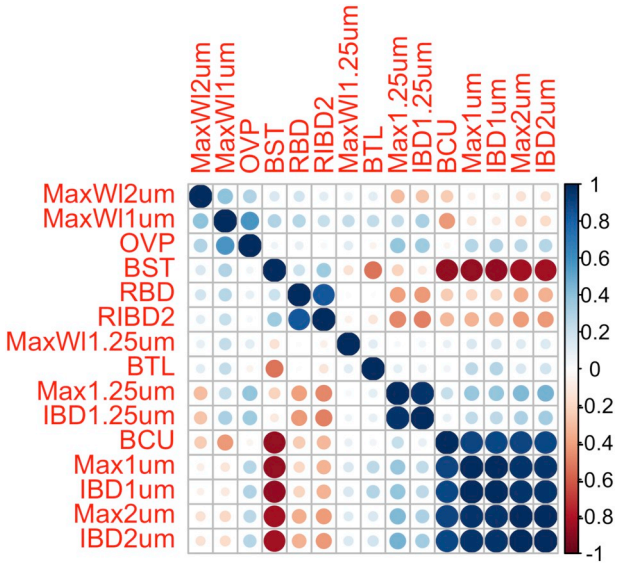
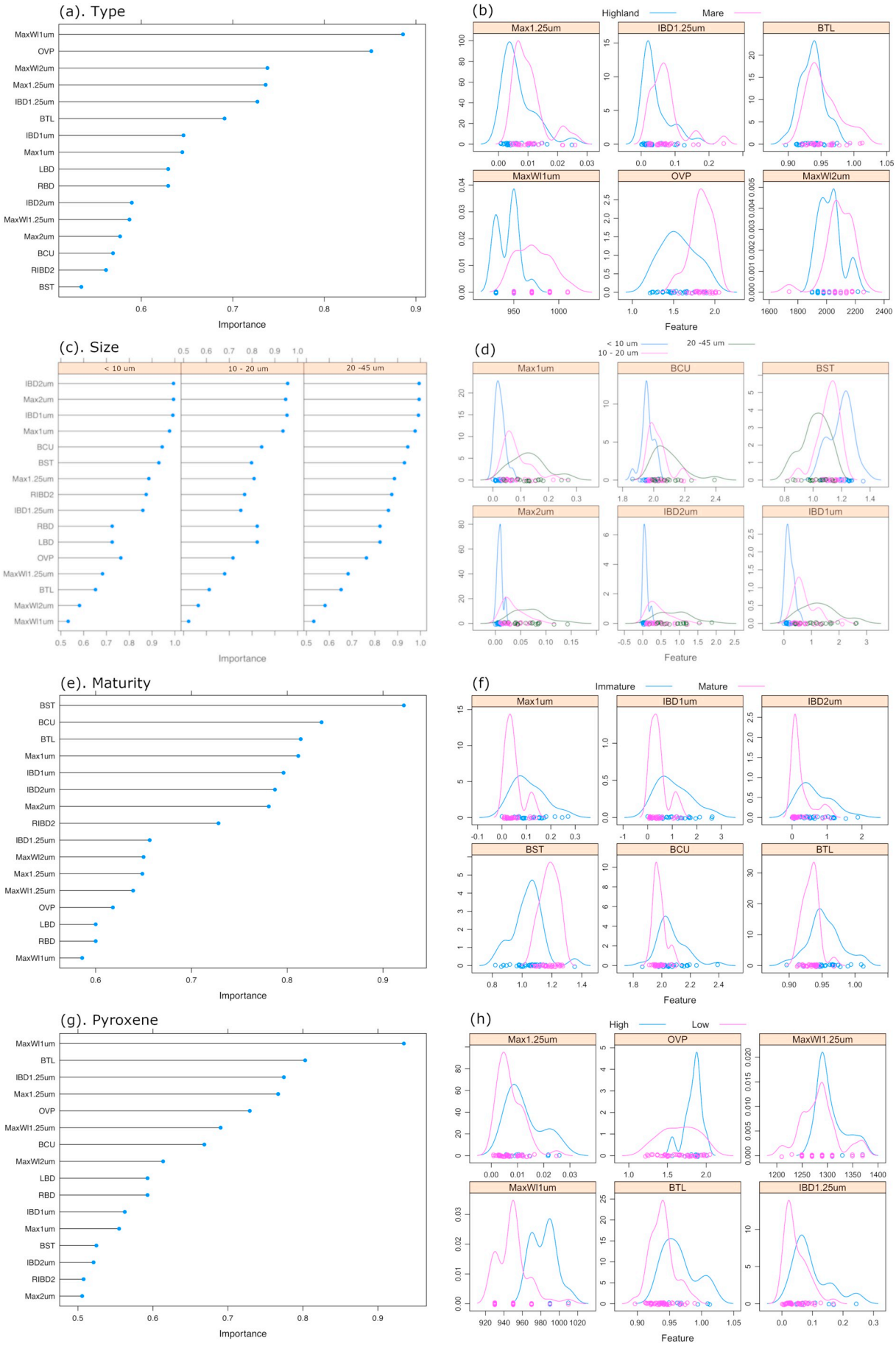


Fig. 5. Correlations between all created features. Dark blue color indicates strong positive correlations, dark red is used for strong negative correlation, and white indicates no empirical relationship between the features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

except for the band curvature (BCU). The BCU shows that, with an increase in grain size, band depths also increase, and values become more dispersed. Grain sizes <10 μm show the lowest and least diverse absorption band depths at 1 μm and 2 μm regardless of their composition. Noble et al. (2006) also showed a strong correlation between the 1 and 2 μm band depths with grain size (except the 10–20 μm fractions of highland soils), but they did not use the <10 μm size fraction for their analysis, because mineral absorption bands in that size faction are too weak to provide reliable information using the MGM model. Band strength (BST in Fig. 6 (d)) decreases with increasing grain size, since the absorption band at 1 μm increases with increasing particle size as the optical path length is shortened (Pieters, 1983). In addition, Fig. 6 (c) shows that the absorption band position at 1 μm and 2 μm is least important, since that band position is not influenced by grain size variations as experimentally confirmed by Crown and Pieters (1987).

However, Band strength was the most important feature for classifying lunar soils based on their maturity, as shown in Fig. 6 (e). The BST feature plot in Fig. 6 (f) also shows clear differences in band strength values between mature and immature soils. Pieters et al. (2001) showed the importance of band strength for mapping maturity in the South Pole-



(caption on next page)

Fig. 6. Feature importance scores and distribution of each attribute by class value for each selected feature. (a) Feature importance for soil type, (b) Conditional density plots of selected features for soil type, (c) Feature importance for size fraction, (d) Conditional density plots of selected features for size fraction, (e) Feature importance for maturity, (f) Conditional density plots of selected features for maturity, (g) Feature importance for type of pyroxene, and (h) Conditional density plots of selected features for the type of pyroxene.

Aitken basin using Clementine data. Band curvature and band tilt also show close relations, as discussed by Pieters et al. (2001). The band position at the 1 μm absorption feature becomes the least important feature for classifying soil maturity because maturity weakens the absorption bands without affecting band positions (Crown and Pieters, 1987; Pieters et al., 2000).

As is well established in the literature (Adams, 1974; Burns, 1970; Crown and Pieters, 1987), the absorption band position at 1 μm is the most important feature for identifying low-Ca and high-Ca pyroxene. Band tilt is the next most important feature and band depth at 2 μm was the least important feature in this case. The MaxWl1um feature plot in Fig. 6 (g) clearly shows two feature classes for low-Ca pyroxene and high-Ca pyroxene. The 1 μm absorption band shifts to longer wavelength with an increase in the Ca content of the pyroxene (Adams, 1974).

2.5. Classification algorithms

A diverse array of ML algorithms including data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide applications in many fields (e.g. science, engineering, medical and business) (Jordan and Mitchell, 2015). ML algorithms can be categorized into three classes; a) linear models that seek to discover the underlying structure of the data using linear combinations of features, b) non-linear models that are highly non-linear functions of the features, and c) classification trees and rule-based models. Classification trees and rule-based models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition, and can be represented graphically as a decision tree. Three ML algorithms from each category were selected for this study, covering a wide range of characteristics such as simplicity, flexibility, computational complexity, and interpretability along with their capability of handling different types of data.

Linear models include Linear Discriminant analysis (*lda*), the Partial Least Squares method (*pls*) and Lasso and the Elastic-Net Regularized Generalized Linear Model (*glmnet*). *lda* is a statistical technique for finding the linear combination of features that best separates observations into different classes. It assumes that the data in each class is normally distributed and that there is a unique covariance matrix for each class. *lda* is sensitive to near zero variance features and collinear features. Therefore, input features should be centered and scaled and near zero variance features should be removed to achieve the correct result (Kuhn and Johnson, 2013). *pls* is a generalization of multiple linear regression methods. *pls* can analyze strongly collinear, correlated, and noisy data. It has been shown to be a highly effective quantitative analysis tool to measure material compositions using ultraviolet, visible and near infrared spectral data (Li, 2006). The *glmnet* module, developed by Friedman et al. (2010), fits a generalized linear model with penalized maximum likelihood. In other words, *glmnet* combines the generalized linear model with elastic net models using an elastic net. The model can work on very large datasets and can take advantage of sparsity in the feature set (Adler, 2012; Friedman et al., 2010).

Non-linear methods include Support Vector Machine (*SVM*), Naïve Bayes, and Feed-Forward Neural Networks with a single hidden layer (*nnet*). SVM was introduced by Boser et al. (1992) to find a linear separating hyperplane (a plane in multidimensional space) that separates classes of interest (Pal and Watanachaturaporn, 2004). In most cases, when the classes in the dataset are mixed, the linear hyperplane cannot separate those classes without misclassification. Therefore, we used the non-linear Radial basis kernel function with SVM (*svmRadial*) as an SVM model. SVMs do not rely on all the underlying data to train the model. They use only some of the observations called support vectors (Boser et al., 1992). Therefore, it is somewhat resistant to outliers. Naïve Bayes is a simple learning algorithm that uses the Bayes rule with a strong assumption that in the given class, the attributes are conditionally independent (Sammut and Webb, 2011). Naïve Bayes always uses all attributes for all features and is therefore relatively insensitive to the noise in the features to be classified. It is also relatively insensitive to missing values in the training data due to its probabilistic nature (Sammut and Webb, 2011).

Neural Networks are learning machines, comprising a large number of neurons which are connected to each other in a layer fashion (Kononenko and Kukar, 2007). Simple neural network contains inputs, outputs and one hidden layer. The number of inputs and output nodes depends on the training set. A small number of hidden layers has two basic advantages, 1) the efficiency of each node increases and therefore the time of computer simulations is significantly reduced, and 2) the network can better generalize the input pattern, resulting in superior predictive power. Therefore, we used a single hidden layer feed-forward neural network. However, using a neural network for classification has a significant potential for overfitting. Additionally, collinearity and non-informative features will have a comparable effect on model performance (Kuhn and Johnson, 2013).

Classification trees and rule-based models include Random Forest (*rf*), C5.0 model, and Boosted Trees (*bstTree*). Random forests are a combination of tree predictors such that each tree in the forest depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). The error rate on the random forest depends on the correlation between any two trees in the forest and the strength of each individual tree in the forest. Increasing the correlation between two trees increases the forest error rate. A tree with a low error rate is a strong classifier. This is an effective method for estimating missing data and it even maintains accuracy when a large proportion of the data is missing. C5.0 is a more advanced version of the tree- and rule-based Quinlan's C4.5 classification model (Kuhn and Johnson, 2013; Sutton, 2005). C5.0 measures the feature importance by determining the percentage of training set samples that fall into each of the terminal nodes after the split. C5.0 also has an option to remove features by identifying the features that do not have a relationship with the outcome. After identifying those non-informative features, C5.0 recreates the tree (Kuhn and Johnson, 2013). A *bstTree* is a method of combining classifiers, which are iteratively created from a weighted version of the learning sample. Weights are adaptively adjusted in each step to give increased weight to the classes which were misclassified in the previous step. The final predictions are obtained by weighting the results of iteratively produced intermediate predictors (Sutton, 2005). Therefore, boosted trees can be used very successfully without fine tuning and tweaking the classifiers (Kuhn and Johnson, 2013; Sutton, 2005). The main difference between the random forest and boosted tress is that in the random forest all trees are created independently having maximum depth (interaction depth) and each tree contributes equally to the final model. But in boosted trees, trees are dependent on past trees having minimum depth and they contribute unequally to the final model (Kononenko and Kukar, 2007; Kuhn and Johnson, 2013).

2.6. Selection of the best model

We use the k-fold CV method to estimate the test error associated with the selected machine learning models to evaluate their performance for each category. The test error is the average error that results

from a machine learning method when it predicts the response to a new observation, which was not used to train the model. k-fold CV involves randomly dividing the set of observations into k groups of approximately equal size. The first group (fold) is treated as a validation set while others (k -1 folds) are used to train the model. The k-fold CV for classification is computed by,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i$$

where Err_i is the number of misclassified observations.

There is a bias-variance trade-off associated with the selection of k in k-fold CV. k-fold cross validation results with k = 5 or k = 10 have been shown empirically to yield test error rate estimates that are not affected by excessively high bias or from very high variance (Kononenko and Kukar, 2007). Therefore, the 10-fold CV method was chosen to validate the modeling results.

We used the training dataset to train the models. To select the best model for each category, five feature combinations were used based on the ranking of feature performance calculated in Section 2.4. The method employed is named the simple backward elimination wrapper method. First, we select the best five feature combinations, and based on the importance value of each feature, we eliminate the least important feature for each instance and then use the selected feature combinations in the model. Table 4 shows the name of the model with associated feature combinations used for each model. The model name consists of a letter followed by a number, where the letter indicates the target variable of the model. If the model is used to classify the samples by Type (mare or highland), it is denoted by T, while S denotes a model used to classify samples based on their size fraction. M denotes the models which used to classify the sample based on target Maturity, and P denotes the target variable Pyroxene (Low-Ca and High-Ca). The number represents the number of features used in each model. Each model in Table 4 was tested with nine machine learning algorithms as discussed in Section 2.5. Each algorithm was repeated three times with 10-fold CV to get a more accurate result (totaling 5400 computations; 4 (category) X 5 (combinations) X 9 (algorithms) X 10 (data splitting) X 3 (repeats)).

3. Results & discussion

3.1. Model performance

Model performance was measured using two statistical measures: overall accuracy and Kappa. The overall accuracy reflects the agreement between the observed and predicted classes. It is simply calculated by the number of all correct predictions divided by the total number of predictors in the dataset (James et al., 2017). The Kappa statistic is a measure of how well the classifier performed as compared to how well it would have performed simply by chance. It is calculated from the observed and expected frequencies on the diagonal of a confusion matrix (Cohen, 1960). A confusion matrix is a simple cross-tabulation of the observed and predicted classes of the model results. The diagonal cells of a confusion matrix indicate the cases where the classes are correctly predicted, and the off diagonals denote the number of errors for each possible case (Cohen, 1960; Kuhn and Johnson, 2013).

$$Kappa = \frac{P_o - P_e}{1 - P_e}$$

where, P_o is the observed accuracy and P_e is the expected accuracy based on the marginal totals of the confusion matix (e. g. Table 5). Kappa values of 1 indicate perfect agreement and a value of zero would indicate a lack of agreement. Less common negative Kappa values indicate a negative association between the observed and predicted data (Kuhn, 2008).

Since each algorithm repeats 30 times for each model, mean accuracy and mean Kappa values with their variance are used to measure the performance of each model. The best model for each feature combination in each category was selected after analyzing the accuracy and Kappa value of each algorithm on each feature combination (Fig. 7). As an example, the highest accuracy with the highest Kappa value was achieved by the Support Vector Machine (*svmRadial*) algorithm with the S2 model combination (Fig. 8).

Each data point in Fig. 7 shows the mean accuracy of each winning algorithm and the model feature combination. The mean Kappa value associated with each model is shown in the "bar" of each point (the performance of every model is presented in the supplementary documents). For each category the best model with its associated ML algorithm was selected based on the mean accuracy and mean Kappa value of each combination and the number of features associated with the model. We have given priority to a smaller number of features in feature combinations to reduce complexity and to increase interpretability. If two models have the same mean accuracy and the same mean Kappa with the same variation, we select the model with the smaller number of features. The target variable Type in Fig. 7 shows that accuracy increases with an increased number of features up to a certain level and then decreases. Therefore, the best feature combination and ML algorithm to identify soil type using spectral features will be the band depths near the 1 μm, 1.25 μm, and 2 μm absorption bands and the olivine/pyroxene

Table 4
Feature combinations used in different models.

Name	MaxWl1um	Max1um	IBD1um	MaxWl1.25um	Max1.25um	IBD1.25um	MaxWl2um	Max2um	IBD2um	BCU	BST	BTL	OVP
T6	x			x		x	x					x	x
T5	x			x		x	x						x
T4	x			x			x						x
T3	x						x						x
T2	x												x
S6		x	x					x	x	x	x		
S5		x	x					x	x	x			
S4		x	x					x	x				
S3			x					x	x				
S2								x	x				
M6		x	x						x	x	x	x	
M5		x	x							x	x	x	
M4		x								x	x	x	
M3										x	x	x	
M2										x	x		
P6	x			x	x	x						x	x
P5	x				x	x						x	x
P4	x				x	x						x	
P3	x					x						x	
P2	x											x	

Table 5
Confusion Matrices of the best models.

a). Type

Prediction	Reference	
	Highland	Mare
Highland	6	1
Mare	0	4
Accuracy	90.9%	
Kappa	81.3%	
Sensitivity	100%	
Specificity	80.0%	

b). Size

Prediction	Reference		
	< 10 μm	10–20 μm	20–45 μm
<10 μm	2	1	0
10–20 μm	1	2	1
20–45 μm	0	0	2
Accuracy	66.6%		
Kappa	50.0%		
Sensitivity	66.6%	66.6%	66.6%
Specificity	83.3%	66.6%	100%

c). Maturity

Prediction	Reference	
	Immature	Mature
Immature	5	1
Mature	1	4
Accuracy	81.8%	
Kappa	63.3%	
Sensitivity	83.3%	
Specificity	80.0%	

d). Pyroxene

Prediction	Reference	
	High-Ca	Low-Ca
High-Ca	2	0
Low-Ca	0	8
Accuracy	100%	
Kappa	100%	
Sensitivity	100%	
Specificity	100%	

ratio with the *glmnet* ML method. In the grain size category, the best feature combination and ML algorithm was the S4 (Band depth and IBD values of 1 μm and 2 μm) and *svmRadial*. Maturity can be correctly classified using the *svmRadial* method with the M4 (Max1um, BCU, BST, and BTL) feature combination. This model shows that MaxWl1um and BTL alone are sufficient to correctly classify the high-Ca pyroxene and low-Ca pyroxene from reflectance spectra. In other words, an increase of feature combinations (after P2) did not increase the model performance.

Errors have varying effects on models' predictive performance and can be introduced in three fundamental ways. a) Errors can be introduced during the calculation of spectral parameters. The accuracy of parameters such as Max1um and IDB1um mainly depend on accuracy of the continuum removal spectra. Those errors are likely to be propagated through the model prediction equation resulting in poor model performance. b) Noise can be introduced into the model by inclusion of non-informative features. Some models can filter out non-informative features, resulting in high predictive performance (Kuhn and Johnson, 2013). For model performance for size fraction (Fig. 7), *glmnet* with four feature combinations (Band depth and IBD values of 1 μm and 2 μm absorption bands) had the highest accuracy. Thus, the inclusion of a new feature (BCU) decreases the accuracy of the model. c) Noise can also be introduced into the model through the response variable, mostly due to the mislabeling of training data. We found that manual binning of continuous I_s/FeO data into three classes significantly decreases the model performance for this reason. Therefore, classification based on the clustering method is adopted here to increase the predictive power of the model.

3.2. Model evaluation

The feature combination with the winning ML algorithm in each category was used to evaluate the predictive power of the model using (unseen) validation data. Table 5 shows the confusion matrices of each selected model from each category, which was evaluated with respective validation datasets. Overall accuracy is not a good measure for the performance of a model if the distribution of the classes is imbalanced, such as in the case of the type of pyroxene in the present dataset. In such cases, one can always select the class having the highest observations and obtain good accuracy performance. Overall accuracy does not make a distinction about the type of error being made. Therefore, sensitivity and specificity measures are also used here to assess the accuracy of the models. Sensitivity measures the rate at which the event of interest is predicted correctly for all samples having the event (i.e. how well the classifier can recognize the positive samples).

$$\text{Sensitivity} = \frac{\#\text{samples with the event and predicted to have the event}}{\#\text{samples having the event.}}$$

Specificity is defined as the rate at which non-event samples are predicted as non-events (i.e. how well the classifier can recognize the negative samples) (James et al., 2017; Kuhn and Johnson, 2013).

$$\text{Specificity} = \frac{\#\text{samples without the event and predicted as nonevent}}{\#\text{samples without the event}}$$

The above confusion matrices indicate that except for size fraction, all categories were classified with >80% accuracies. All models achieved 80% or higher Sensitivity and Specificity, except for the target variable Size. The Support Vector Machine with radial basis function correctly classified the maturity into two classes except for two observations. Since SVM is a black-box model, it is difficult to learn anything about a problem by looking at the parameters from a fitted SVM model. Those two points might represent the fuzzy boundary between two

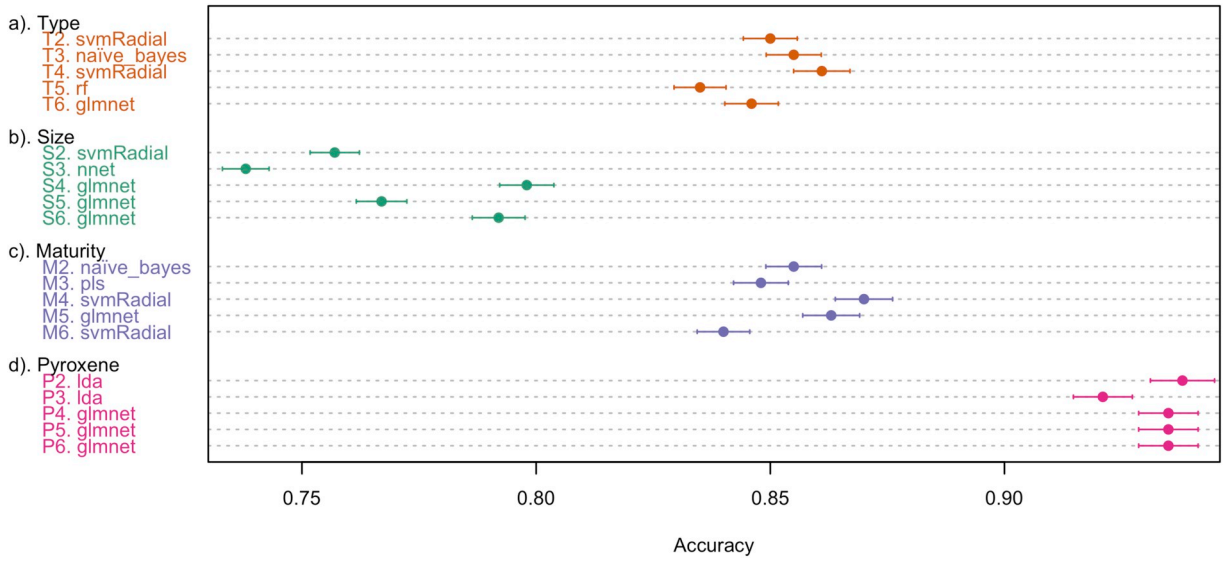


Fig. 7. Performance of the best model for each feature combination.

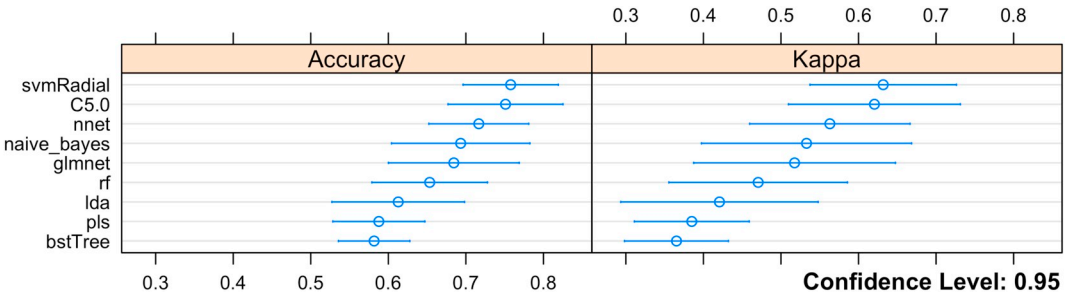


Fig. 8. Classification accuracy of target variable Size using Max2um and IBD2um.

classes. Classification of pyroxene using linear discriminant analysis with the features of 1 μm band position and band tilt (BTL) yielded 100% accuracy.

4. Future work

Findings of this research will be used to map the lunar surface based on the type of soil (highland and mare), maturity and type of dominant pyroxene (high Ca and low Ca) using Moon Mineralogy Mapper hyperspectral data.

5. Conclusions

Nine different ML algorithms representing a wide range of inherent capabilities were applied to the spectral parameters derived from LSCC data to predict several physical and mineralogical properties of lunar soils from their reflectance spectra. Here we summarize several important results:

1. Band depths at 1 μm, 1.25 μm and 2 μm with band index OVP (R_{2018}/R_{1008}) was the best feature combination to distinguish the highland and mare soils using their reflectance spectra. Non-linear SVM with radial basis kernel function was able to classify these successfully with 90% accuracy. The 1.25 μm absorption band appears to be caused by pyroxene rather than plagioclase.
2. Identifying the soil size fraction using reflectance spectra was difficult. Previous studies also addressed this issue and found that maturity and mineral mixtures have a larger effect on the spectrum than the particle size (e.g. Crown and Pieters, 1987). Band depth and Integrated band depth at 1 μm and 2 μm absorption band provided the best results.
3. Soil maturity can be correctly identified by a feature combination of band depth at 1 μm, band curvature, band strength, and band tilt. A proper binning method is essential for higher accuracy.
4. The feature combination of band position at 1 μm with band tilt was able to distinguish samples with high- or low-Ca pyroxene with 100% accuracy.
5. The Support Vector Machine with radial basis function (*svmRadial*) and penalized logistic regression model (*glmnet*) performed well for all target variables with high accuracies.
6. Since there is an exponential number of potential feature combinations, it is usually not possible to search all possible combinations (2^n possible combinations for n number of features). Therefore, we used a simple backward elimination wrapper approach starting with the full set of features and checked the performance by removing features one at a time. Further study can reveal the best two-feature combination using a heuristic search, since we identified the best four features for each case starting with fifteen features. We also found the best two ML algorithms to analyze spectral parameters of the LSCC data set staring from nine ML algorithms.

Acknowledgements

Many thanks to the R Core Team, and Rstudio for making them as free and open source, all the R library developers for their effort and contribution, and for LSCC, for making it accessible.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.icarus.2020.113719.

References

Adams, J.B., 1974. Visible and near-infrared diffuse reflectance spectra of pyroxenes as applied to remote sensing of solid objects in the solar system. J. Geophys. Res. 79 (32), 4829–4836.
Adams, J.B., 1975. Interpretation of visible and near-infrared diffuse reflectance spectra of pyroxenes and other rock-forming minerals. In: Karr, C. (Ed.), Infrared and Raman Spectroscopy of Lunar and Terestrial Minerals. Academic Press, pp. 91–116.
Adams, J.B., Goullaud, L.H., 1978. Plagioclase feldspars: visible and near infrared diffuse refectance spectra as applied to remote sensing. In: Proc. Lunar Planet. Sci. Conf. 9th, pp. 2901–2909.
Adler, J., 2012. R in a Nutshell. A Desktop Quick Reference. O'Reilly (699 pp).
Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271.
Borst, A.M., Foing, B.H., Davies, G.R., Westrenen, W.v., 2012. Surface mineralogy and stratigraphy of the lunar South Pole-Aitken basin determined from Clementine UV/VIS and NIR data. Planetary and Space Science 68, 76–85.
Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh, PA, pp. 144–152.
Bowles, M., 2015. Machine Learning in Python. Essential Techniques for Predictive Analysis. Wiley (326 pp).
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Burns, R.G., 1970. Mineralogical Applications of Crystal Field Theory. Cambridge Topics in Mineral Physics and Chemistry 5. Cambridge University Press (551 pp).
Cahill, J.T.S., Lucey, P.G., Stockstill-Cahill, K.R., Hawke, B.R., 2010. Radiative transfer modeling of near-infrared reflectance of lunar highland and mare soils. J. Geophys. Res. 115 (E12013) https://doi.org/10.1029/2009JE003500.
Cheek, L.C., et al., 2011. Goldschmidt crater and the Moon's north polar region: results from the moon mineralogy mapper (M3). J. Geophys. Res. 116 (E00G02) https://doi.org/10.1029/2010JE003702.
Cheek, L.C., Hanna, K.L.D., Pieters, C.M., Head, J.W., Whitten, J.L., 2013. The distribution and purity of anorthosite across the Orientale basin: new perspectives from moon mineralogy mapper data. J. Geophys. Res. Planets 118, 1805–1820.
Clark, R.N., 1999. Spectroscopy of Rocks and Minerals, and Principles of Spectroscopy, Manual of Remote Sensing. U.S. Geological Survey, MS 964 Box 25046 Federal Center Denver, CO 80225–0046, pp. 3–58.
Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J. Geophys. Res. 89 (B7), 6329–6340.
Clark, R.N., King, T.V.V., Gorelick, N.S., 1987. Automatic continuum analysis of reflectance spectra. In: Proceedings of the Third Airborne Imaging Spectrometer Data Analysis Workshop, pp. 138–142.
Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G.A., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. 95 (B8), 12,653–12,680.
Cloutis, E.A., Gaffey, M.J., Jackowski, T.L., Reed, K.L., 1986. Calibrations of phase abundance, composition, and particle size distribution for olivine-orthopyroxene mixtures from reflectance spectra. J. Appl. Geophys. 91 (B11), 11641–11653.
Cochran, W.G., 1977. Sampling Techniques. John Wiley & Sons (428 pp).
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. XX (1), 37–46.
Conel, J.E., Nash, D.B., 1970. Spectral reflectance and albedo of Apollo 11 lunar samples: effects of irradiation and vitrification and comparison with telescopic observations. In: Proc. of the Apollo 11 Lunar Science Conference, pp. 2013–2023.

Cook, R.D., 1977. Detection of influential observation in linear regression. Technometrics 19 (1), 15–18.
Crown, D.A., Pieters, C.M., 1987. Spectral properties of plagioclase and pyroxene mixtures and the interpretation of lunar soil spectra. Icarus 72, 492–506.
Denevi, B.W., Lucey, P.G., Sherman, S.B., 2008. Radiative transfer modeling of near-infrared spectra of lunar mare soils: theory and measurement. J. Geophys. Res. 113 (E02003) https://doi.org/10.1029/2007JE002929.
Dhingra, D., 2008. Exploring links between crater floor mineralogy and layered lunar crust. Adv. Space Res. 42, 275–280.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 22.
Green, A.A., Craig, M.D., 1985. Analysis of aircraft spectrometer data with logarithmic residuals. In: Vane, G., Goetz, A. (Eds.), Proceedings of the Airborne Imaging Spectrometer Data Analysis Workshop, vols. 86-35. JPL Publication, pp. 111–119.
Hapke, B., 1993. Theory of reflectance and emittance spectroscopy. In: Topics in Remote Sensing, 3. Cambridge University Press (455 pp).
Hareyama, M., et al., 2019. Global classification of lunar reflectance spectra obtained by Kaguya (SELENE): implication for hidden basaltic materials. Icarus 321, 407–425.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2017. An introduction to statistical learning with applications in R. In: Springer Texts in Statistics. Springer (426 pp).
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260.
Kodikara, G.R.L., Champati ray, P.K., Chauhan, P., Chatterjee, R.S., 2016. Spectral mapping of morphological features on the moon with MGM and SAM. Int. J. Appl. Earth Obs. Geoinf. 44, 31–41.
Kohonen, T., 2001. Self-Organizing Maps. In: Springer Series in Information Sciences, 30. Springer (501 pp).
Kononenko, I., Kukar, M., 2007. Machine Learning and Data Mining: Introduction to Principles and Algorithms. Horwood Publishing (454 pp).
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 26.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer (600 pp).
Lehnert, L.W., Meyera, H., Meyerb, N., Reudenbacha, C., Bendix, J., 2014. A hyperspectral indicator system for rangeland degradation on the Tibetan plateau: a case study towards spaceborne monitoring. Ecol. Indic. 39, 54–64.
Lehnert, L.W., et al., 2015. Retrieval of grassland plant coverage on the Tibetan Plateau based on a multi-scale, multi-sensor and multi-method approach. Remote Sens. Environ. 164, 197–207.
Lehnert, L.W., et al., 2018. Hyperspectral data analysis in R: the hsdar package. J. Stat. Softw. 89 (12), 1–23. https://doi.org/10.18637/jss.v089.i12.
Li, L., 2006. Partial least squares modeling to quantify lunar soil composition with hyperspectral reflectance measurements. J. Geophys. Res. 111 (E04002) https://doi.org/10.1029/2005JE002598.
Li, S., Li, L., 2011. Radiative transfer modeling for quantifying lunar surface minerals, particle size, and submicroscopic metallic Fe. J. Geophys. Res. 116 (E09001) https://doi.org/10.1029/2011JE003837.
Li, S., Li, L., Milliken, R., Song, K., 2012. Hybridization of partial least squares and neural network models for quantifying lunar surface minerals. Icarus 221, 208–225.
Liu, D., Li, L., Sun, Y., 2015. An improved radiative transfer model for estimating mineral abundance of immature and mature lunar soils. Icarus 253, 40–50.
Lundeen, S., McLaughlin, S., Alanis, R., 2010. Moon Mineralogy Mapper: Data Product Software Interface Specification, JPL D-39032, Version 9.5.
McCord, T.B., Clark, R.N., Hawke, B.R., McFadden, L.A., Owensby, P.D., 1981. Moon: near-infrared spectral reflectance, a first good look. J. Geophys. Res. 86 (B11), 10883–10892.
McCraig, M.A., et al., 2017. Fitting the curve in Excel®: systematic curve fitting of laboratory and remotely sensed planetary spectra (review article). Comput. Geosci. 100, 103–114.
Morris, R.V., 1976. Surface exposure indices of lunar soils: a comparative FMR study. In: Proceedings of 7th Lunar Science Conference, pp. 315–335.
Morris, R.V., 1978. The surface exposure (maturity) of lunar soils: some concepts and Is/FeO compilation. In: Proc. Lunar Planet. Sci. Conf. 9th, pp. 2287–2297.
Mouelic, S.L., Langevin, Y., 2001. The olivine at the lunar crater Copernicus as seen by Clementine NIR data. Planetary and Space Science 49, 65–70.
Mustard, J.F., Pieters, C.M., 1989. Photometric phase functions of common geologic minerals and applications to quantitative analysis of mineral mixture reflectance spectra. J. Geophys. Res. 94 (B10), 13619–13634.
Mustard, J.F., et al., 2011. Compositional diversity and geologic insights of the Aristarchus crater from Moon Mineralogy Mapper data. J. Geophys. Res. 116 (E00G12) https://doi.org/10.1029/2010JE003726.
Nash, D.B., Conel, J.E., 1974. Spectral reflectance systematics for mixtures of powdered hypersthene, labradorite, and ilmenite. J. Geophys. Res. 79 (11), 1615–1621.
Noble, S.K., Pieters, C.M., Hiroi, T., Taylor, L.A., 2006. Using the modified Gaussian model to extract quantitative data from lunar soils. J. Geophys. Res. 111 (E11009) https://doi.org/10.1029/2006JE002721.
Pal, M., Watanachaturaporn, P., 2004. Support vector machines. In: Varshney, P.K., Arora, M.K. (Eds.), Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data. Springer, pp. 133–157.
Pieters, C.M., 1983. Strength of mineral absorption features in the transmitted component of near-infrared reflected light: first results from RELAB. J. Geophys. Res. 88 (B11), 9534–9544.
Pieters, C.M., Englert, P.A.J., 1993. Remote geochemical analysis: elemental and mineralogical composition. In: Topics in Remote Sensing. Cambridge University Press (594 pp).
Pieters, C.M., et al., 2000. Space weathering on airless bodies: resolving a mystery with lunar samples. Meteorit. Planet. Sci. 35, 1101–1107.
Pieters, C.M., James, W., Head, I., Gaddis, L., Jolliff, B., Duke, M., 2001. Rock types of south pole-Aitken basin and extent of basaltic volcanism. J. Geophys. Res. 106 (E11), 28,001–28,022.
Pieters, C.M., Stankevich, D.G., Shkuratov, Y.G., Taylor, L.A., 2002. Statistical analysis of the links among lunar mare soil mineralogy, chemistry, and reflectance spectra. Icarus 155, 285–298.
Pieters, C.M., Shkuratov, Y., Kaydash, V., Stankevich, D., Taylor, L., 2006. Lunar soil characterization consortium analyses: pyroxene and maturity estimates derived from Clementine image data. Icarus 184, 83–101.
Reitermanova, Z., 2010. Data Splitting, WDS' 10 Proceedings of Contributed Papers. MATFYZPRESS, pp. 31–36.
Sammut, C., Webb, G.I. (Eds.), 2011. Encyclopedia of Machine Learning. Springer (1031 pp).
Sivakumar, V., Neelakantan, R., Santosh, M., 2017. Lunar surface mineralogy using hyperspectral data: implications for primordial crust in the Earth-Moon system. Geosci. Front. 8, 457–465.
Sunshine, J.M., Pieters, C.M., 1993. Estimating modal abundances from the spectra of natural and laboratory pyroxene mixtures using the modified Gaussian model. J. Geophys. Res. 98 (E5), 9075–9087.
Sunshine, J.M., Pieters, C.M., Pratt, S.F., 1990. Deconvolution of mineral absorption bands: an improved approach. J. Geophys. Res. 95 (B5), 6955–6966.
Sutton, C.D., 2005. Classification and regression trees, bagging, and boosting. In: Handbook of Statistics. Elsevier, pp. 303–329.
Taylor, L.A., Pieters, C.M., Keller, L.P., Morris, R.V., McKay, D.S., 2001. Lunar mare soils: space weathering and the major effects of surface-correlated nanophase Fe. J. Geophys. Res. 106 (E11), 27,985–27,999.
Taylor, L.A., et al., 2010. Mineralogical and chemical characterization of lunar highland soils: insights into the space weathering of soils on airless bodies. J. Geophys. Res. 115 (E02002), E02002 https://doi.org/10.1029/2009JE003427.
Tompkins, S., Pieters, C.M., 1999. Mineralogy of the lunar crust: results from Clementine. Meteorit. Planet. Sci. 34, 25–41.
Tompkins, S., Pieters, C.M., Mustard, J.F., Pinet, P., Chevrel, S.D., 1994. Distribution of materials excavated by the lunar crater Bullialdus and implications for the geologic history of the Nubium region. Icarus 110, 261–274.
Trang, D., Lucey, P.G., 2019. Improved space weathering maps of the lunar surface through radiative transfer modeling of Kaguya multiband imager data. Icarus 321, 307–323.
Ward, J.H., 1963. Hierarchical grouping to optimize and objective function. J. Am. Stat. Assoc. 58 (301), 236–244.
Wenxiang, X., et al., 2019. New maps of lunar surface chemistry. Icarus 321, 200–215.