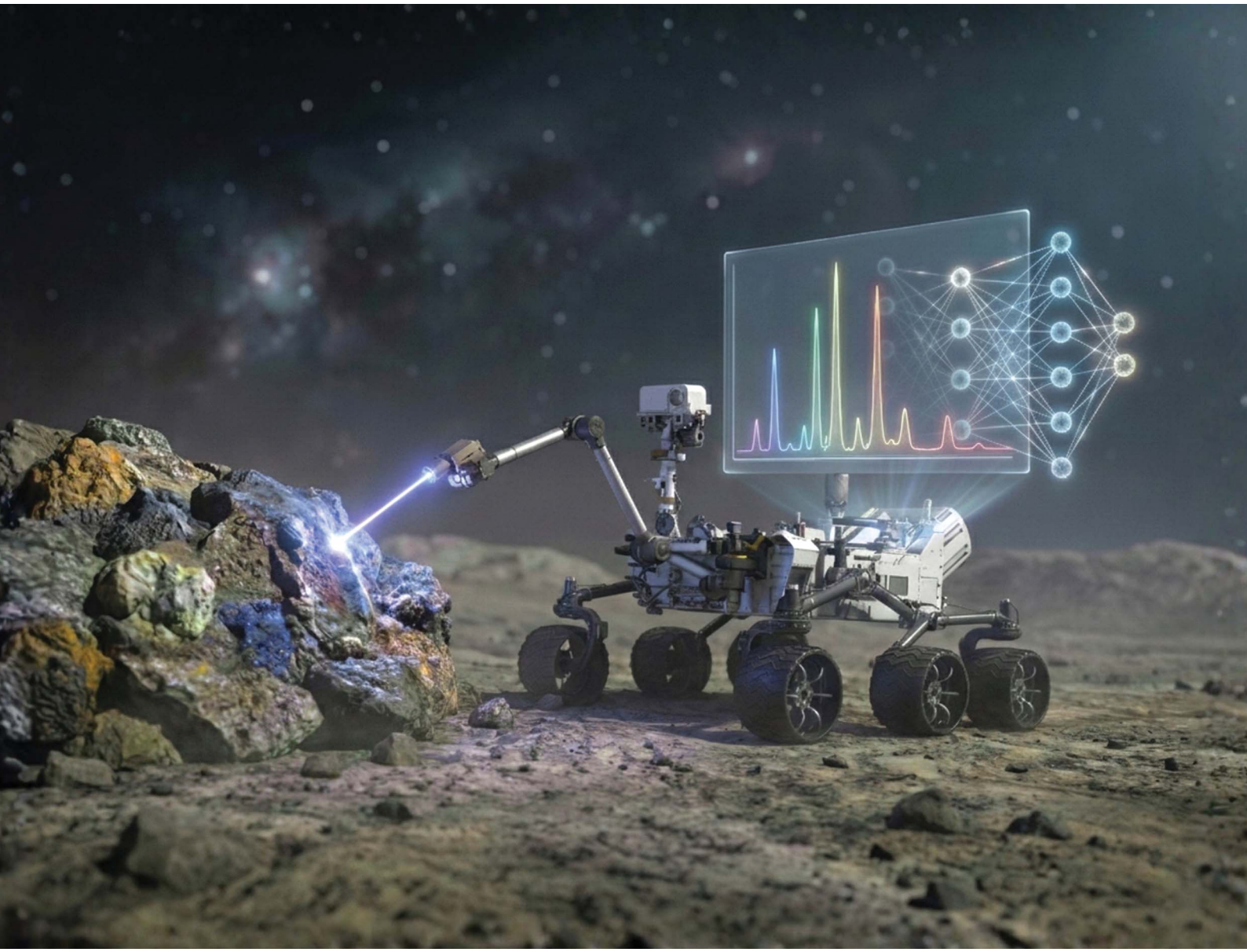
Volume 41
Number 2
February 2026
Pages 537–806

JAAS

Journal of Analytical Atomic Spectrometry

rsc.li/jaas



ISSN 0267-9477

ROYAL SOCIETY OF CHEMISTRY

PAPER

Homa Saeidfirozeh *et al.*

Towards advanced mineral identification for future space mining applications employing LIBS and machine learning

JAAS

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *J. Anal. At. Spectrom.*, 2026, **41**, 615

Towards advanced mineral identification for future space mining applications employing LIBS and machine learning

Homa Saeidfirozeh, *[a] Ashwin Kumar Myakalwar, [b] Pavlína Šeborová, [a] Ján Žabka,[a] Bernd Abel, [ac] Petr Kubelík[a] and Martin Ferus [a]

The growing interest in sustainable space exploration has brought *in situ* resource utilization (ISRU) to the forefront of planetary science. This study presents an integrated approach to autonomous mineral identification for space mining by combining Laser-Induced Breakdown Spectroscopy (LIBS) with supervised machine learning (ML). A dataset of over 400 high-resolution LIBS spectra representing 25 mineral classes was collected under simulated low-pressure conditions to replicate extraterrestrial environments. The raw spectra were preprocessed using wavelet-based denoising to reduce random noise, baseline correction to remove the background continuum, and spectral normalization to account for intensity variations. To simplify the data and enhance classification performance, three feature selection methods were applied: Principal Component Analysis (PCA), which identifies directions of maximum variance to reduce data dimensionality; variance thresholding, which removes spectral features with negligible variability across samples; and random forest-based feature selection (RF-FS), which ranks wavelengths by their importance for classification. Several classification algorithms were evaluated, with test accuracies reaching up to 89.3%. The best results were achieved using random forest and logistic regression models trained on features selected by RF-FS, showing strong generalization to previously unseen samples. This work demonstrates the potential of LIBS-ML integration for fast, robust, and accurate mineral classification, including reliable identification of dominant phases in mineral mixtures in planetary environments. The approach also provides interpretability and classifier confidence estimation, supporting adaptive autonomous mineral identification for future robotic exploration missions.

Received 28th September 2025
Accepted 25th November 2025

DOI: 10.1039/d5ja00377f

rsc.li/jaas

1 Introduction

Human relationship with celestial treasures dates back to the dawn of civilization. Ancient societies integrated the night sky into their myths, rituals, and material culture, often attributing divine significance to meteoritic materials. A striking example is Pharaoh Tutankhamun's iron dagger, forged from meteoritic metal.[1] In a broader sense, the Sun has served as the earliest and most enduring extraterrestrial resource, providing light, warmth, and the foundation of agriculture. In modern times, this natural inheritance has evolved into the strategic use of solar energy in space missions.[2] These historical layers illustrate how our engagement with the cosmos has gradually transformed from symbolic reverence to the practical pursuit of off-Earth resources.

By the late 19th century, the idea of utilizing resources from beyond Earth began to appear in speculative literature. One notable early example is Garrett P. Serviss's Edison's Conquest of Mars (1898), which describes the extraction of gold from asteroids, foreshadowing modern concepts of asteroid mining. By the 20th century, forward thinkers began imagining how we might utilize resources from beyond our planet. Konstantin Tsiolkovsky, a pioneering rocket scientist, set the groundwork for space travel and dreamed of humans expanding into space, supported by materials found on other worlds.[3] Afterwards, in the mid-1900s, innovators like Arthur C. Clarke played a key role in popularizing these concepts by introducing them to the public through science fiction and futuristic ideas, inspiring many to imagine mining the Moon and asteroids,[4] and laying the foundation for today's serious discussions about space resource utilization (SRU).[5] These concepts evolved into institutional strategies emphasizing the importance of *in situ* resource utilization (ISRU), reducing reliance on Earth-supplied materials, and enabling affordable long-term space missions.[6]

[a]*J. Heyrovský Institute of Physical Chemistry, Czech Academy of Sciences, Dolejškova 3, CZ 18223 Prague 8, Czech Republic. E-mail: homa.saeidfirouzeh@jh-inst.cas.cz*
[b]*Department of Physics, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation Higher Education, Dontanpally, Hyderabad 501203, India*
[c]*Faculty of Chemistry and Mineralogy, Institute of Chemical Technology, Leipzig University, Linnéstraße 3, 04103 Leipzig, Germany*

Defining "space resources" has become increasingly critical: a material qualifies if it is present in a useful concentration, extractable with foreseeable technology, and serves practical space operations or markets. Recent efforts are underway to adapt terrestrial mineral classification standards, such as the Lunar Ore Reserves Standard (LORS-101),[7] explicitly designed to categorize extraterrestrial deposits by feasibility and utility. Over recent decades, space resource mapping through remote sensing and sample analysis has progressed significantly.[5] Agencies have identified promising concentrations of elements such as Fe, Ti, and Si, and extensive deposits of water ice in the Moon's polar regions, considered to be important for future propellant production.[8–10] Despite promising orbital and remote sensing data, the actual composition, spatial distribution, and accessibility of these extraterrestrial resources remain uncertain. Reliable *in situ* measurements are therefore crucial for validating resource models and developing effective extraction strategies.[11] Laser-Induced Breakdown Spectroscopy (LIBS) has emerged as a powerful tool for real-time geochemical analysis in space exploration.[12] LIBS enables direct analysis of unprepreprocessed, unpolished surfaces with almost any geometry, making it highly suitable for elemental characterization in extraterrestrial environments. While not directly involved in material extraction, LIBS plays a critical role in resource prospecting and compositional mapping, foundational steps toward the realization of extraterrestrial mining. Unlike traditional methods, LIBS operates effectively in low gravity and vacuum, providing real-time elemental composition analysis without the need for extensive sample preparation or complex instrumentation, making it a suitable analytical technique for different planetary missions.[13–15] Despite these advantages, interpreting LIBS spectra under field conditions is challenging due to spectral complexity. Here, recent advances in Machine Learning (ML) offer a transformative approach. ML algorithms trained on known spectral signatures enable efficient classification of minerals from noisy or novel spectra, significantly improving speed and accuracy *in situ*. On Earth, LIBS has proven its versatility in analyzing diverse mineral and ore samples, including pyrite (FeS_2), hematite (Fe_2O_3), molybdenite (MoS_2), and chalcopyrite ($CuFeS_2$), many containing economically valuable metals like copper, iron, zinc, and tungsten. This study features a mineral campaign including copper ores (azurite and malachite), iron ores (hematite and magnetite), and rarer materials like bauxite (aluminum ore) and wolframite (tungsten ore), simulating the diversity expected in extraterrestrial mining environments.

However, LIBS alone cannot efficiently handle the vast and complex spectral datasets generated during extraterrestrial mining. This challenge aligns with the broader scientific priorities highlighted by the Mars Sample Return initiative, which emphasizes the necessity for precise, rapid, and robust analytical techniques to exploit returned planetary samples[6,16] fully. In this study, we focus on the ultraviolet (UV) spectral region under atmospheres of 10 and 10^{-2} mbar, closely simulating the low-pressure conditions encountered in space. ML provides a transformative solution, enabling the accurate classification and analysis of diverse mineral samples in real-time. Unlike deep learning methods that require large datasets, ML algorithms such as Random Forest (RF), Support Vector Machines (SVM), *K*-Nearest Neighbors (KNN), and Logistic Regression (LR) excel in small-data environments typical of space missions. These approaches effectively manage sparse, imbalanced data and noisy spectra, making them invaluable for on-the-fly decision-making in space resource extraction. Moreover, ML approaches are not limited to mineral classification; recent studies have successfully applied neural networks to predict plasma parameters directly from LIBS spectra, such as plasma temperature estimation using synthetic ChemCam-based simulations,[17] and rapid detection of trace elements like xenon in complex plasma mixtures relevant for geochemical and planetary analyses.[18]

This study presents a robust LIBS-ML integration methodology that bridges Earth-based experiments with extraterrestrial resource exploration, addressing challenges like matrix effects, spectral noise, and small dataset variability. A key distinguishing feature of this work is its focus on mineral identification using LIBS spectra collected under planetary-like low-pressure conditions, as well as a careful evaluation of performance on complex mineral mixtures, which closely reflects the real-world scenarios of future space resource utilisation.

The paper is organized as follows. Sect. Materials and Methods, details the experimental methods and describes data processing and ML; sect. Results and Discussion presents the results and discussion; the next section discusses strategies to handle novel data and improve autonomous decision making in remote applications, and the last section concludes with key findings and future outlook.

2 Materials and methods

2.1 Experimental

2.1.1 Materials. Certified reference mineral samples used in this study were purchased from a well-established Czech mineral supplier.[29] Table 1 lists the ores and rock-forming minerals, featuring economically important materials such as hematite, magnetite, bauxite, and cassiterite, which are key resources for industrial and space mining. Additionally, rock-forming minerals, including olivine, feldspar, gypsum, serpentinite, and dolomite, were included due to their known occurrence in Martian geology,[21] emphasizing their relevance to Martian *in situ* resource utilization (ISRU). We carefully selected samples to cover a wide variety of minerals, like silicates, oxides, sulfides, carbonates, and native elements, mirroring the kind of mineral diversity typically found on planetary surfaces.

Fig. 1 shows the surface of the bauxite sample as an example. The experiment was performed on this surface to enhance laser absorption, improve plasma formation, and reduce reflectivity and matrix effects.[30] The dark lines visible in the marked area indicate the effect of laser ablation, where material removal and surface modification have occurred due to the interaction of the laser with the sample. To capture a representative analysis, multiple laser spots were applied across a broader area, capturing different microstructures within the sample matrix.

Table 1 Ores and rock-forming minerals were analyzed in this study for LIBS-based classification, focusing on Martian resource utilization

Material name	Metal/Material mined	Chemical formula	Mineral classification	Geological occurrence	Spectroscopic signatures	Ref.
Ores						
Azurite	Copper	$Cu_3(CO_3)_2(OH)_2$	Carbonate	Oxidized zones of copper deposits	Distinct Cu peaks at 324.8 nm and 327.4 nm	—
Bauxite	Aluminum	$Al(OH)_3$	Hydroxide	Lateritic soils in tropical regions	Broad Al peaks around 394.4 nm and 396.1 nm	—
Bismuth	Bismuth	Bi	Native element	Hydrothermal veins	Bi spectral line at 306.7 nm	—
Cassiterite	Tin	SnO_2	Oxide	Igneous and metamorphic rocks	Sn lines at 189.9 nm and 317.5 nm	—
Chalcopyrite	Copper	$CuFeS_2$	Sulfide	Hydrothermal veins, igneous rocks	Cu peaks at 324.8 nm and 327.4 nm	19
Chalcocite	Copper	Cu_2S	Sulfide	Supergene enrichment zones	Cu peak at 324.8 nm	—
Chromite	Chromium	$FeCr_2O_4$	Oxide	Ultramafic rocks	Cr peaks around 425.4 nm	20
Kyanite	Aluminum	Al_2SiO_5	Silicate	Metamorphic rocks	Al peak at 396.1 nm	—
Galena	Lead	PbS	Sulfide	Hydrothermal veins	Pb peak at 220.3 nm	—
Goethite	Iron	FeO(OH)	Hydroxide	Secondary mineral in iron deposits	Fe peaks at 259.9 nm and 271.9 nm	21
Grossular	Aluminum	$Ca_3Al_2(SiO_4)_3$	Silicate	Metamorphic rocks	Al peaks at 394.4 nm and 396.1 nm	—
Hematite	Iron	Fe_2O_3	Oxide	Sedimentary and metamorphic rocks	Fe peaks at 259.9 nm and 372.0 nm	21
Magnetite	Iron	Fe_3O_4	Oxide	Igneous and metamorphic rocks	Fe peak at 516.7 nm	22
Malachite	Copper	$Cu_2CO_3(OH)_2$	Carbonate	Oxidized zones of copper deposits	Cu peaks at 324.8 nm and 327.4 nm	23
Molybdenite	Molybdenum	MoS_2	Sulfide	Hydrothermal veins	Mo peaks at 390.3 nm and 386.4 nm	—
Pyrite	Sulfur	FeS_2	Sulfide	Sedimentary and hydrothermal deposits	Fe peaks at 259.9 nm and 371.9 nm	24
Sphalerite	Zinc	ZnS	Sulfide	Hydrothermal veins	Zn peak at 213.8 nm	—
Stibnite	Antimony	Sb_2S_3	Sulfide	Hydrothermal veins	Strong Sb lines in the UV range	—
Wolframite	Tungsten	$(Fe,Mn)WO_4$	Tungstate	Hydrothermal veins	W peaks at 207.9 nm and 255.2 nm	—
Zircon	Zirconium	$ZrSiO_4$	Silicate	Igneous and metamorphic rocks	Zr peaks at 343.8 nm and 349.6 nm	25
Rock-forming minerals						
Olivine	Magnesium, iron	$(Mg,Fe)_2SiO_4$	Silicate	Ultramafic rocks (peridotites, basalts)	Fe peaks at 516.7 nm, Mg lines in UV.	26
Gypsum	Calcium	$CaSO_4 \cdot 2H_2O$	Sulfate	Sedimentary deposits, evaporites	Ca peaks at 393.3 nm, 396.8 nm	27
Feldspar	Aluminum, K, Na	$(K,Na,Ca)AlSi_3O_8$	Silicate	Igneous/metamorphic rocks	Al peaks at 394.4 nm, 396.1 nm	21
Serpentine	Magnesium	$(Mg,Fe)_3Si_2O_5(OH)_4$	Silicate	Metamorphic rocks (alteration of peridotite)	Mg peaks in UV, Fe peaks at 259.9 nm	21
Dolomite	Magnesium, calcium	$CaMg(CO_3)_2$	Carbonate	Sedimentary rocks, hydrothermal veins	Ca peaks at 393.3 nm, Mg peaks in UV.	28

2.1.2 Data acquisition and initial conditions. The samples were ablated using a Nd : YAG laser (Nd-doped yttrium-aluminum garnet laser), which provides a pulse with a wavelength of 1064 nm, a duration of 6 ns, and an energy of 450 mJ. The repetition rate was set to 10 Hz. A CaF_2 lens with a focal length of 10 cm was used to focus the laser beam on the samples attached to a moving stage located inside a vacuum chamber. The measurements were carried out under two different pressure conditions: 10 mbar and 10^{-2} mbar.

Emission spectra of the laser ablation-induced plasma were recorded using the high-resolution Butterfly Echelle spectrograph, equipped with an Andor ICCD camera. The spectrograph operates in the UV (192–433 nm) region, offering spectral resolutions of 13–31 pm (with a resolving power of 14 000). The Echelle spectrograph was set to trigger 50 ns after the laser pulse and collect the signal for 1 μs. The final spectra were obtained by accumulating 20 laser shots. Before data collection, the Butterfly spectrograph was calibrated using a Hg lamp.

Fig. 1 Top view of a bauxite sample as an example of a mineral used in LIBS analysis. The white rectangular box marks the area where the laser spots were applied during measurements.

2.2 Computational method

2.2.1 Data processing of LIBS spectra. Twenty-five different ore and mineral families were examined to design our dataset, maximizing the mineral diversity essential for utilizing space resources by including many mineral families, each characterized by a smaller number of carefully selected high-quality spectra. This approach aligns with previous studies that emphasize broad coverage of composition through representation of diverse classes.[31]

As the first step, careful filtering criteria were applied to improve LIBS data quality. These criteria focused on analyzing the standard deviation and noise distribution to exclude spectra with excessive noise or poorly defined peaks, which ensured that only high-quality spectra were included in the dataset. Of the 437 spectra across 25 ore and mineral families, 417 spectra fulfilled these criteria, resulting in a retention rate of 95.5%, which confirms that most of the data were preserved for analysis. A filtering efficiency of 4.5% represents the proportion of spectra removed due to low quality, demonstrating the effectiveness of the filtering process. Additional details on the preprocessing steps, including examples of excluded spectra with high noise or anomalously intense peaks, are provided in the SI.

Then, we analyzed the noise distribution across all spectra by calculating the standard deviation of the first 50 intensity points for each spectrum. As shown in Fig. 2, the normalized histogram of noise levels follows a Gaussian distribution, with most spectra exhibiting noise levels ranging between 1.1 and 2.3. The noise levels were normalized to a density, ensuring comparability across datasets and enabling a fitted normal distribution overlay. The mean noise level was calculated to be $\mu = 1.64$, with a standard deviation of $\sigma = 0.17$, indicating that the noise distribution is tightly clustered around the mean. Spectra with noise levels below 1.06 or above 2.32 were rare, demonstrating the uniform quality of the dataset. A threshold of 1.49, derived as the 20th percentile of the noise levels, was used to identify spectra with low noise for subsequent analysis, ensuring robust preprocessing and consistent data quality.

As the next step after noise analysis, we implemented wavelet denoising to improve spectral data quality by reducing noise while maintaining important signal features.[32] This process involves three key steps. First, the signal $x(t)$ is decomposed into wavelet coefficients $c_{i,j}$ and wavelet basis functions $\psi_{i,j}(t)$ using the Daubechies-4 (db4) wavelet:

$$x(t) = \sum_{i=1}^{N} \sum_{j=1}^{2^i} c_{i,j}\psi_{i,j}(t) \tag{1}$$

Next, a soft thresholding function is applied to the wavelet coefficients to suppress noise while maintaining the significant components of the signal. The soft thresholding function used is defined as:[33]

$$\hat{c}_{i,j} = \begin{cases} \operatorname{sign}(c_{i,j})\left(|c_{i,j}| - \lambda\right), & \text{if } |c_{i,j}| > \lambda \\ 0, & \text{if } |c_{i,j}| \le \lambda \end{cases} \tag{2}$$

Here, λ represents the threshold value, derived from the noise analysis results. Finally, the denoised signal $\hat{x}(t)$ is reconstructed by applying the inverse wavelet transform to the thresholded coefficients:

$$\hat{x}(t) = \sum_{i=1}^{N} \sum_{j=1}^{2^i} \hat{c}_{i,j}\psi_{i,j}(t) \tag{3}$$

This denoising approach was implemented using the PyWavelets library,[34] with soft thresholding explicitly applied through the pywt.threshold function. The Daubechies-4 wavelet was selected for its optimal balance between resolution and smoothness, making it well-suited for LIBS spectral data.[32] This process enhanced spectral quality by approximately 1.5 times, underscoring its effectiveness in improving data reliability and enabling more accurate downstream analyses (see the example in the SI).

For baseline correction, the process uses wavelet decomposition to separate the spectrum $y(\lambda)$ into a low-frequency

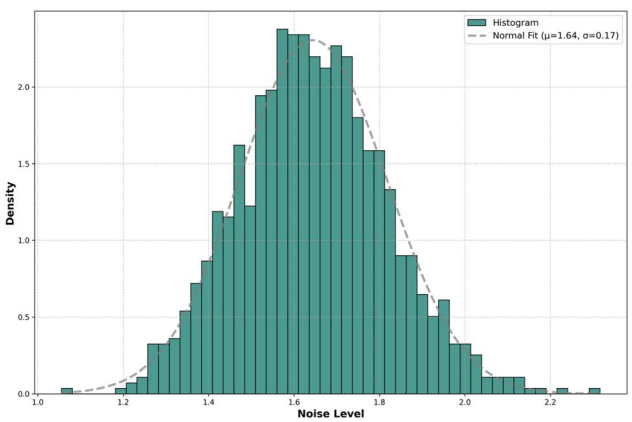


Fig. 2 The histogram shows the noise level distribution across spectra, while the grey dashed line represents the fitted normal distribution.

baseline $A_j(\lambda)$ and high-frequency details $D_i(\lambda)$, isolating the baseline by setting $D_i(\lambda) = 0$. The baseline is adjusted to ensure $A_j(\lambda) \leq y(\lambda)$, preventing overcorrection. The corrected spectrum is then calculated as $y_{\text{corrected}}(\lambda) = y(\lambda) - A_j(\lambda)$. This approach effectively removes the baseline while preserving the spectral peaks, ensuring no negative values or artificial elevation in the corrected spectrum.

Then, each spectrum x_i was normalized using:

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j} \tag{4}$$

where $x_{i,j}$ is the intensity at wavelength j for sample i, $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j}$ is the mean, and σ_j is the standard deviation for feature j across all samples.

It is worth noting that across all tested minerals, the spectra collected at different pressures showed strong correlations, with only minor intensity variations and no significant peak shifts or formation of new lines. Therefore, standard normalization and preprocessing are sufficient to combine or compare data at both pressures for mineral classification.

2.2.2 Machine learning approach. Afterwards, the dataset was randomly divided into training and test sets using a stratified sampling approach.[35] This process ensures that each class is represented in the training and test sets in the same proportions as in the original data. In other words, for a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with K classes, stratified sampling aims to maintain the same probability distribution $p(y = k)$ in both the training and test groups.

For hyperparameter optimization and model selection, we further applied k-fold cross-validation within each training set.

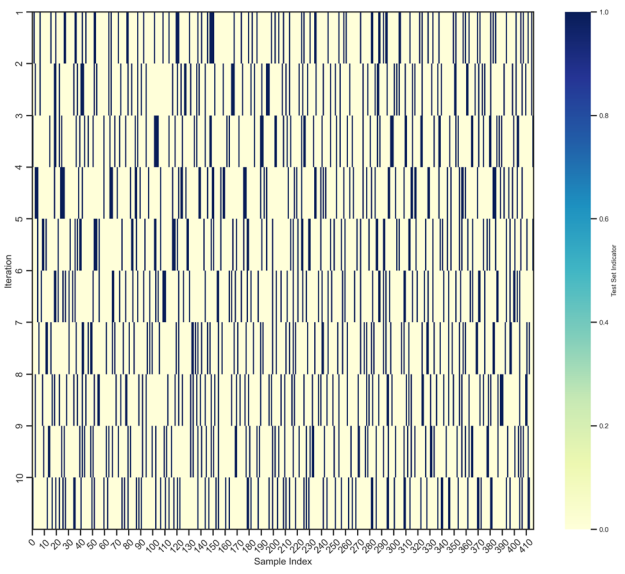


Fig. 3 The heatmap illustrates how our dataset was split for model evaluation across 10 different random iterations. Each row represents a different split, and each column is a sample from the dataset. The blue cells mark which samples were selected as part of the test set for that particular iteration. By repeating this process 10 times, we ensured that every sample had a chance to be tested, providing a fair and thorough assessment of our models while minimising sampling bias.

In this procedure, the training data are partitioned into k equally sized folds; each fold serves as a temporary validation set while the remaining $k - 1$ folds are used for training. This process is repeated k times, allowing every sample to be used for validation exactly once. The model's performance is then averaged across all k folds, yielding a robust and unbiased estimate of generalization accuracy. In this study, we used $k = 5$.

Fig. 3 shows the results of repeated random stratified splits: each row is an iteration, each column a sample, and blue cells mark test set assignments. This demonstrates that every sample is included in the test set throughout the cross-validation process.

To address the issue of class imbalance in our dataset, we applied the Synthetic Minority Oversampling Technique (SMOTE),[36] which generates synthetic minority class samples by interpolating between existing minority instances:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \delta \times (\mathbf{x}_{\text{nn}} - \mathbf{x}_i), \tag{5}$$

where $\mathbf{x}_i$ is a minority class sample, $\mathbf{x}_{\text{nn}}$ is one of its nearest neighbors, and $\delta \in [0, 1]$ is a random scalar. SMOTE was applied to the training set before feature selection and classifier training, so that all models benefit from a balanced class distribution during fitting.

To improve robustness, we performed repeated random splits and implemented k-fold cross-validation, which provides a more reliable estimate of model performance. In k-fold cross-validation, the data are partitioned into k equal-sized subsets $(\mathcal{D}_1, \ldots, \mathcal{D}_k)$. For each fold j, the model is trained on $\mathcal{D}/\mathcal{D}_j$ and tested on $\mathcal{D}_j$, cycling through all k folds. The average performance is calculated as:

$$\text{CV}_{\text{score}} = \frac{1}{k}\sum_{j=1}^{k} \text{score}_j \tag{6}$$

where k is the number of folds, and score_j is the evaluation metric on the j-th fold.[37] Before training the classifiers, we explored three different feature selection and dimensionality reduction approaches, which are described in the next section.

2.2.3 Feature selection and dimensionality reduction methods

2.2.3.1 Principal component analysis. Principal Component Analysis (PCA) is a widely used unsupervised dimensionality reduction technique in LIBS data analysis. By projecting the original data onto a new set of orthogonal axes (principal components), PCA captures the directions of maximum variance. Mathematically, PCA solves the eigenvalue problem for the covariance matrix of the data:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{X} \tag{7}$$

where $\mathbf{X}$ is the mean-centered data matrix, and $\mathbf{C}$ is the covariance matrix. The principal components are defined by the eigenvectors $\mathbf{w}_k$ of $\mathbf{C}$, and the projection onto each component is calculated as follows:

$$z_k = \mathbf{X}\mathbf{w}_k \tag{8}$$

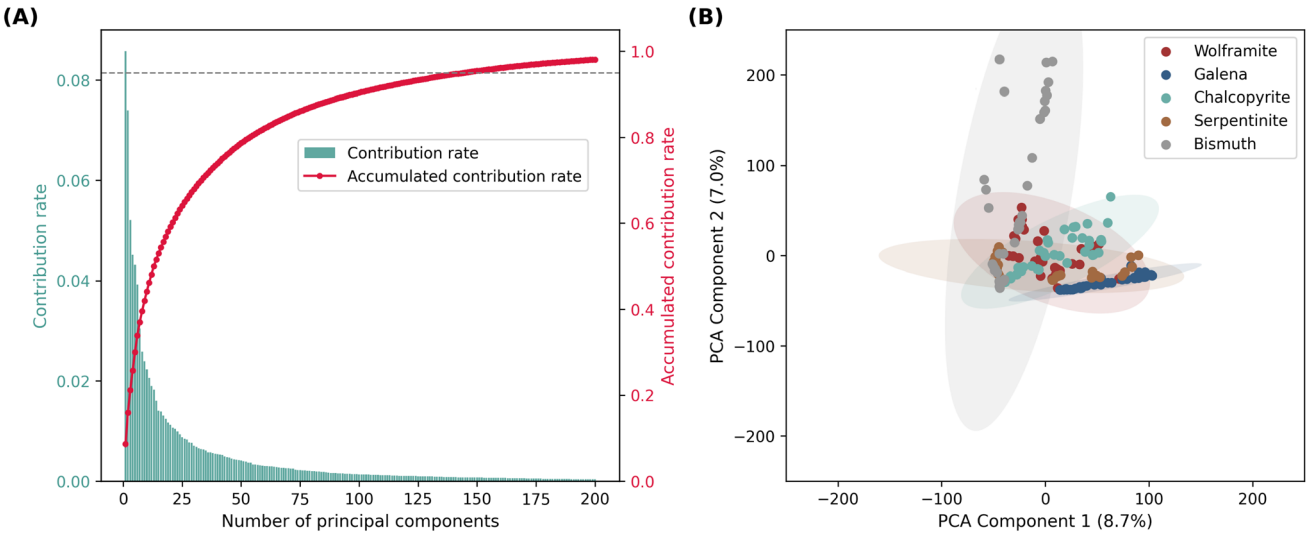


Fig. 4 (A) Number of principal components (*x*-axis) *vs.* contribution rate (left *y*-axis, green bars) and accumulated contribution rate (right *y*-axis, red curve) for the full LIBS dataset. The grey dashed line marks the largest individual contribution rate, where the first principal component alone explains about 8.5% of the total variance. This figure illustrates that most variance is captured by the first few components, supporting the dimensionality reduction in our analysis. (B) Projection of sample spectra from five selected mineral classes onto the first two principal components (PC1 and PC2). Ellipses indicate 95% confidence intervals for each class, illustrating how PCA captures class-specific variance and enables partial separation of mineral families in the reduced-dimensional space.

Fig. 4A shows the distribution of explained variance across the principal components extracted from the LIBS data. The bars on the left axis indicate the individual contribution rate, that is, the fraction of total variance explained by each principal component, corresponding to the eigenvalues introduced above. The red curve displays the accumulated contribution rate as more components are included. As observed, the cumulative explained variance increases steeply, with the curve approaching 1.0 (100%) after relatively few components. This demonstrates that most of the spectral information in the LIBS data can be efficiently captured with a limited number of principal components. Rather than retaining all components, we focus on those that collectively account for at least 95% of the total variance, as this threshold effectively preserves the essential structure and chemical information present in LIBS spectra.[31] To highlight the class differentiation achieved by these principal components, Fig. 4B projects LIBS spectra from five representative mineral classes onto the first two components (PC1 and PC2). The 95% confidence ellipses highlight distinct clustering, indicating that the variance captured in Fig. 4A translates into meaningful spectral separation. While the ellipses represent the main class clusters, the presence of points outside these boundaries is consistent with expected measurement variability and spectral complexity. Together, these panels confirm that PCA not only efficiently reduces dimensionality but also preserves class-specific information critical for accurate classification.

2.2.3.2 Variance threshold. Variance Threshold (VT) is an unsupervised feature selection technique that removes features exhibiting low variance across all samples, assuming that features with very little variation are unlikely to be informative. Given a data matrix **X**, each feature j is retained if its variance satisfies

$$\mathrm{Var}(X_j) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2 > \theta \tag{9}$$

where θ is a predefined threshold, this approach efficiently reduces dimensionality by discarding nearly-constant features without reference to any class labels.

2.2.3.3 Random forest feature selection. Random Forest feature selection (RF-FS) is a supervised approach that leverages the strengths of random forests to identify which features in the data are most effective in distinguishing between classes. As the model learns, it scores each feature by measuring how much it helps reduce uncertainty (or impurity) in the classification process, averaged over all the trees in the forest:

$$I(j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S_{t,j}}\frac{N_s}{N}\cdot\Delta i(s) \tag{10}$$

where I_j is the importance of feature j, T is the number of trees, and $\Delta\mathrm{Impurity}_{j,t}$ is the decrease in impurity caused by feature j in tree t. In the end, we keep the features with the highest scores, focusing the next steps of our analysis on the parts of the data that matter most for predicting the sample's class.

A significant advantage of our RF-FS approach is the interpretability it offers, as it identifies the most important spectral features used for classification. Fig. 5 provides a detailed validation of the RF-FS method used to identify the most informative wavelengths in LIBS spectral data. Fig. 5A illustrates the excellent match between the wavelengths selected by RF-FS and the known atomic emission lines from the NIST database,[38] demonstrating that the model prioritises physically meaningful spectral features.

Fig. 5B shows the selected wavelengths overlaid on the average LIBS spectrum from all samples and mineral families.

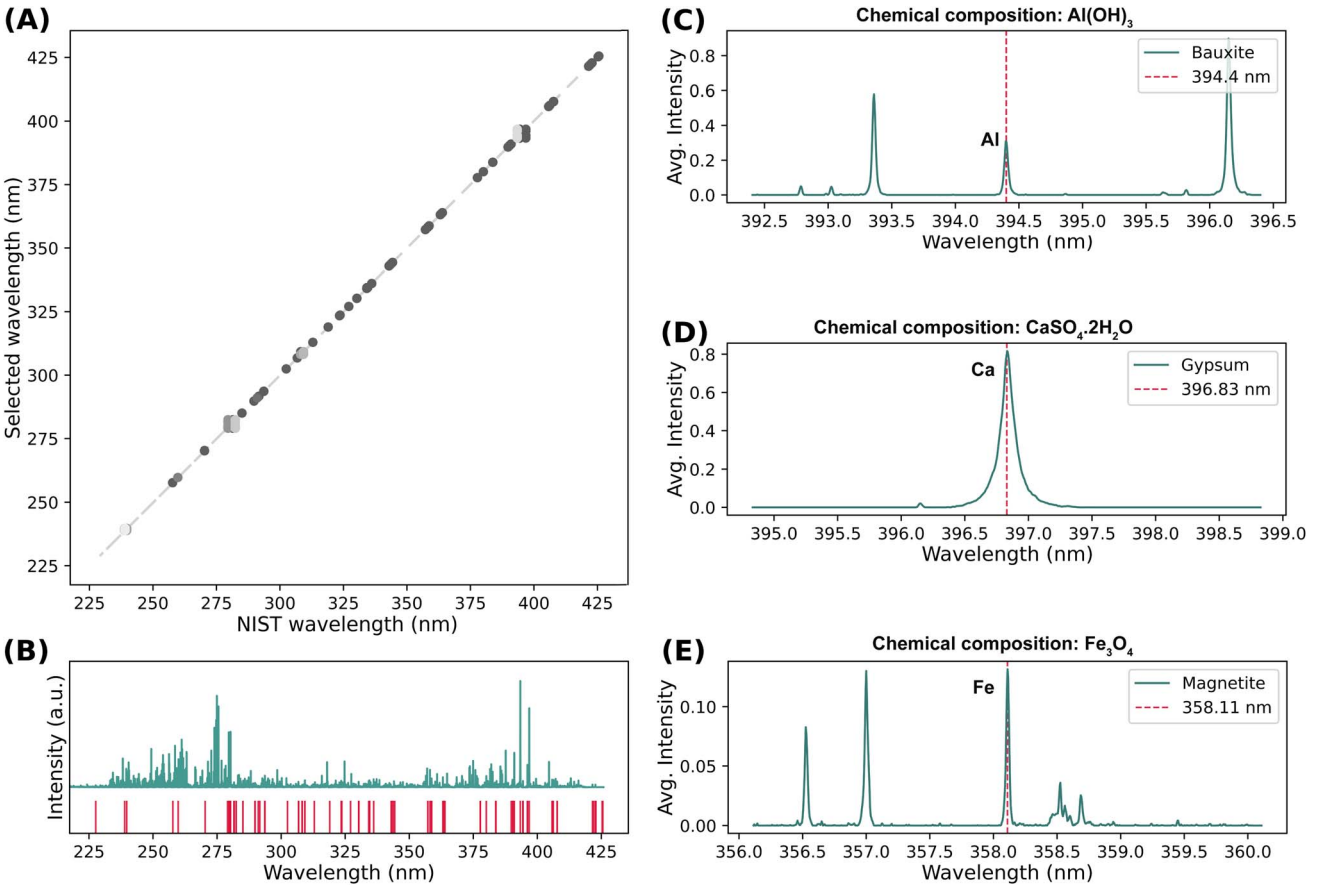


Fig. 5 Random forest feature selection results: (A) selected wavelengths *versus* NIST emission lines, (B) averaged spectrum with selected features, and (C–E) zoomed spectral regions for bauxite, gypsum, and magnetite, highlighting key model-selected emission lines.

This illustrates how the RF-FS method effectively focuses on the prominent spectral peaks, which help distinguish between different minerals. Fig. 5C presents a zoomed spectrum of the averaged bauxite sample; as indicated in Table 1, bauxite is rich in Al, and the wavelength selected by the model at 394.40 nm corresponds to one of the most prominent Al emission lines. Meanwhile, the second line at 396.15 nm is also visible; both lines are key spectral markers widely used for aluminum detection in spectroscopic analysis.[39] Moreover, panel D zooms in on gypsum, highlighting the important 396.83 nm Ca emission line identified by the model, which has been shown to reliably correlate with calcium concentration variations under different experimental conditions in remote LIBS analysis.[40] Panel E showcases magnetite with the key Fe line at 358.11 nm.[38] Moreover, as our spectral window is limited to the UV region, some of the most intense sodium (Na) and potassium (K) emission lines, such as the prominent Na doublet at 589 nm, fall outside the measured range. Nevertheless, the RF-FS method consistently selected alternative Na emission features present within the UV window, as detailed in the SI. The RF-FS method, applied across all tested classifiers, identifies the most important spectral features aligned with known elemental lines, thereby improving classification accuracy while making the models more interpretable.

2.2.4 Classification models. The features obtained from the previous dimensionality reduction and selection steps were then used as input for four supervised classification models, as described below. These models were selected for their complementary strengths in handling multiclass, high-dimensional LIBS data and for their well-established use in spectroscopy-based classification tasks.[41]

2.2.4.1 Random forest. This model is an ensemble of decision trees, where the final prediction is based on majority voting:

$$\hat{y} = \text{mode}(\{T_1(x), T_2(x), \ldots, T_m(x)\}) \quad (11)$$

where $T_i(x)$ is the prediction of the i-th tree.[42] The main hyperparameters tuned for RF are the number of trees in the ensemble ($n_{\text{estimators}}$) and the proportion of features considered at each split ($\max_{\text{features}}$), both of which control the model's complexity and diversity.

2.2.4.2 Support vector machine. This model finds the hyperplane maximizing class separation. The decision function is calculated as follows:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \quad (12)$$

where $K(x_i, x)$ is the kernel function, and α_i and b are model parameters.[43] The main hyperparameters for SVM are the choice of the kernel (linear, polynomial, RBF, or sigmoid) and the regularization parameter (C), which controls the trade-off between maximizing margin and minimizing classification error.

2.2.4.3 *K*-nearest neighbors. This model classifies a sample based on the majority class of its k nearest neighbors:

$$\hat{y} = \text{mode}(\{y^{(1)}, y^{(2)},\ldots,y^{(k)}\}) \tag{13}$$

where $y^{(i)}$ is the class of the i-th nearest neighbor.[44] The main hyperparameter is the number of neighbors (k) used for voting. In this study, we also varied the number of principal components retained after PCA as an additional parameter during the grid search.

2.2.4.4 Logistic regression. This models class probabilities using the sigmoid function as follows:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x+b)}} \tag{14}$$

where w and b are the weights and bias, respectively.[45] LR was tuned for the regularization penalty type (L_1 or L_2) and regularization strength (C), both of which help prevent overfitting by shrinking model coefficients.

These methods were chosen for their effectiveness and diversity in handling classification tasks. A detailed analysis and comparison of classifier performance are presented in the following section.

2.2.5 Classifier performance evaluation. For all classifiers, sample labels were assigned according to the class with the highest predicted probability (argmax rule), so that each sample was always classified. No fixed probability cutoff was used to abstain or flag uncertain predictions. However, classifier confidence (predicted probability or entropy) was calculated for each prediction, and ROC curves were generated by varying the decision threshold over the full probability range. We measured classifier performance using the following standard metrics.

2.2.5.1 Accuracy (Acc). The proportion of correctly classified samples is calculated as follows:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i) \tag{15}$$

where N is the total number of samples, y_i is the true label, $\hat{y}_i$ is the predicted label, and $\mathbb{I}$ is the indicator function.

2.2.5.2 F1 score (F1). The harmonic mean of precision and recall (reported as a weighted average for multiclass) is represented as follows:

$$F_1 = 2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

See the SI for details.

2.3.5.3 Balanced accuracy (BAcc). The average recall across all classes is calculated as follows:

$$\text{BAcc} = \frac{1}{K}\sum_{k=1}^{K}\frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k} \tag{17}$$

where K is the number of classes, and TP_k, FN_k are true positives and false negatives for class k.

2.3.5.4 Matthews correlation coefficient (MCC).

$$\text{MCC} = \frac{\text{TP}\cdot\text{TN} - \text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{18}$$

where TP, TN, FP, FN are total true/false positives/negatives. For multiclass problems, MCC is generalised as described in ref. 46.

2.2.5.5 Cross-validation variance (σ_{CV}^2). The variance of validation accuracy across k folds is calculated as follows:

$$\sigma_{\text{CV}}{}^2 = \frac{1}{k}\sum_{j=1}^{k}(a_j - \bar{a})^2 \tag{19}$$

where a_j is accuracy in fold j, and $\bar{a}$ is the mean accuracy across folds.

These metrics were computed for each classifier and feature selection combination to provide a robust and multi-faceted evaluation of model performance. The results of these analyses, including detailed grid search optimization, classifier comparisons, and performance summaries, are presented in the following section.

3 Results and discussion

After applying the preprocessing steps to the raw data described in Sec. 2.2.1, the spectral data were ready for ML. Each spectrum was modeled as a high-dimensional feature vector made up of intensity values measured over a consistent range of wavelengths. These intensity values reflect the presence and relative abundance of specific elements in the sample, as revealed by the characteristic emission lines in the LIBS spectra. The consistent wavelength grid ensured that the intensity values across all spectra were dimensionally aligned, enabling direct comparison and analysis. The high-resolution Butterfly Echelle spectrograph used in this study produces over 47 000 variables per spectrum, corresponding to data points across the wavelength range of 187.19 to 425.85 nm. The dataset consists of 417 spectra across 25 different classes, forming a large data matrix of 417 rows (one for each spectrum) and 47 693 columns (each representing intensity at a specific wavelength). Handling such high-dimensional data can be computationally challenging. That's why efficient dimensionality reduction, along with careful preprocessing and noise reduction, is crucial for removing redundant information and enhancing the effectiveness of the analysis. In this matrix, each row captures the intensity profile of a single spectrum, while each column corresponds to the intensity at a particular wavelength. This organized format was then used as input for ML classifiers, which learned to recognize patterns and differences between the classes. By using intensity values as features, the models could directly extract meaningful insights from the LIBS spectral data, leading to reliable and accurate classification. As mentioned earlier, to address class imbalance, we employed the SMOTE technique.

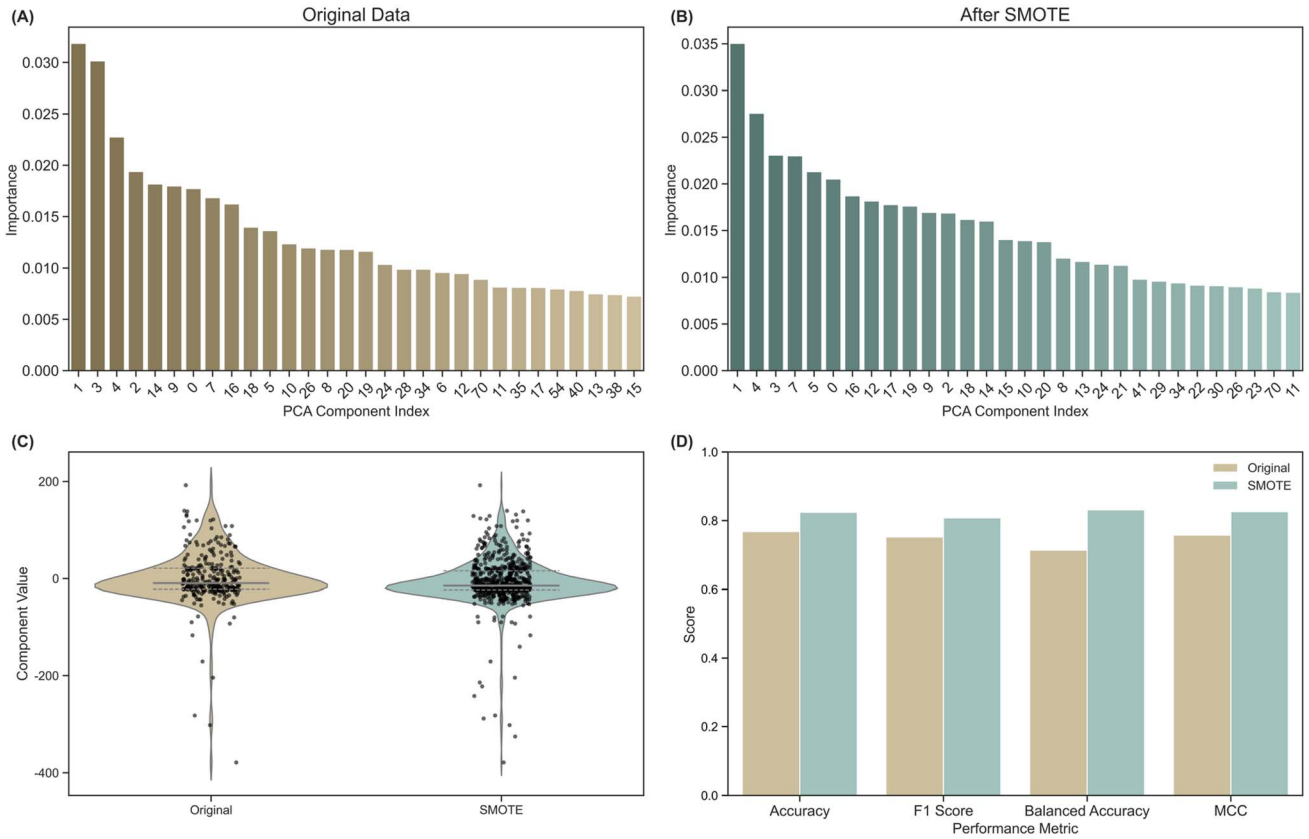


Fig. 6 (A) Principal component importance before SMOTE; (B) importance after SMOTE. (C) Violin plots of the first principal component values for each sample before and after SMOTE; each dot represents a sample's score on PC1, and the width of the violin at any value reflects the density of samples (wider regions = more samples). This highlights reduced skewness and fewer extreme values after balancing. (D) Summary of classifier metrics, demonstrating improved accuracy, F1, balanced accuracy, and MCC with SMOTE.

Fig. 6 shows how SMOTE affects the RF model when using PCA for feature selection. Fig. 6A and B show the importance of principal components before and after applying SMOTE, highlighting how balancing the dataset shifts the relevance of different components. This indicates that SMOTE changes the underlying data structure, influencing which features are most informative for classification. Fig. 6C compares the distribution of the first principal component before and after applying SMOTE. Each dot in the violin plot represents the PC1 value for a single sample; the width of the violin at any point reflects how many samples have similar PC1 values (*i.e.*, wider areas indicate higher sample density). In this context, skewness refers to the asymmetry of the distribution, with longer tails indicating more samples with extreme values. Before SMOTE, the distribution is wider with noticeable extreme values, or long tails, on both sides, indicating that the data are more spread out and skewed. This skewness reflects the imbalance and variability in the original dataset. After SMOTE, the distribution becomes narrower and more symmetric, with fewer extreme values and reduced skewness. This change suggests that SMOTE has effectively balanced the dataset by mitigating extreme variability and bias toward outlying values, leading to a more representative and stable feature distribution. Although the Kolmogorov–Smirnov test[47] showed a non-significant statistical difference ($p = 0.150$), the observational evidence supports the positive impact of SMOTE on data balance. Finally, D presents a comparison of performance metrics, demonstrating clear improvements after applying SMOTE. The next step was to fine-tune the hyperparameters of each classifier to achieve the best possible predictive performance. This was done using a comprehensive grid search, a well-established method that systematically tests different model settings. The results are shown in Fig. 7, which displays grid search heatmaps for all four classifiers paired with PCA-based feature selection. Similar grid searches were performed for RF-based and VT feature selection; however, only the PCA results are presented here for clarity.

Fig. 7A shows how RF accuracy varies depending on the number of trees used in the ensemble (n_estimators) and the fraction of features considered at each decision point (max_features). The model performs best, often with test accuracy above 0.95, when both the number of trees and the feature proportion are set high, suggesting that a larger, more diverse ensemble leads to stronger and more reliable classification. This finding is in agreement with Sheng *et al.*, who reported near-perfect classification accuracy for iron ore samples by optimizing these parameters in their RF models.[48] Moreover, this improvement can be explained by the theoretical generalization error bound of RF introduced by Breiman:[42]

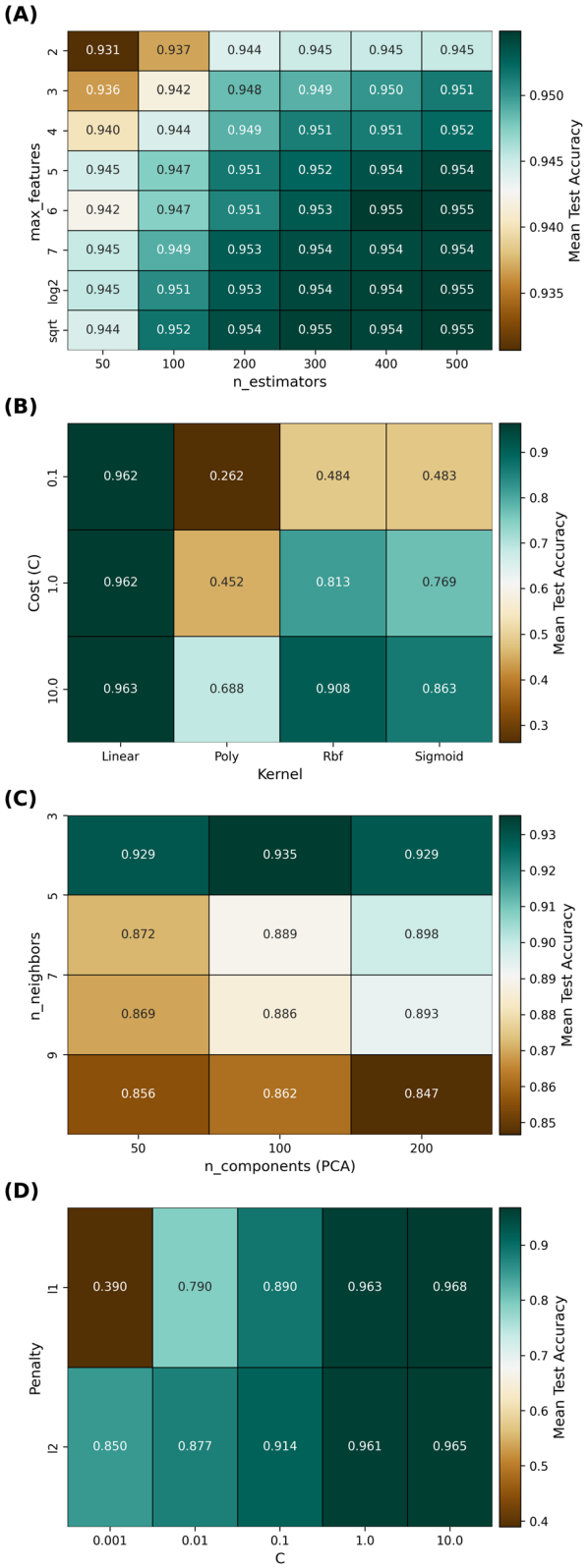


Fig. 7 Grid search hyperparameter optimization results for four classifiers applied to LIBS-based mineral identification: (A) RF (n_estimators and max_features), (B) SVM (kernel and C), (C) KNN (n_neighbors and PCA components), and (D) LR (regularization strength C and penalty type). Values indicate mean test accuracy across parameter grids.

$$\mathrm{PE} \leq \rho \frac{1 - s^2}{s^2} \tag{20}$$

where PE is the prediction error, s is the strength of individual trees, and ρ is the average correlation between trees. Increasing $n_{\text{estimators}}$ stabilizes the ensemble by averaging many trees, while increasing $\max_{\text{features}}$ reduces correlation ρ by introducing randomness. Together, these reduce the overall error, improving classification accuracy.

Fig. 7B shows the classification accuracy of SVM models across different kernels and regularisation parameters (C). The linear kernel consistently achieves the highest accuracy around 0.96, across all tested C values, indicating that the LIBS spectral data are largely linearly separable in the original feature space. This is expected since LIBS spectra often contain prominent, distinctive peaks corresponding to elemental signatures, which can be effectively separated using linear decision boundaries. The RBF kernel yields moderately high accuracy (up to 0.91) but exhibits more variability, depending on parameter settings, suggesting a limited non-linear structure. The sigmoid kernel exhibits intermediate performance, while the polynomial kernel performs the worst, especially at lower C values, likely due to overfitting or a mismatch in model complexity with the data characteristics. Additional hyperparameter tuning results across different kernels and parameter grids are available in the SI.

Fig. 7C highlights the dependence of KNN classification accuracy on the number of neighbors (k) and the number of principal components retained after PCA. The best accuracy (≈ 0.94) is achieved with three neighbors and 100 principal components, suggesting that an optimal balance between dimensionality reduction and neighborhood size improves performance. The KNN classifier predicts the class of a sample based on majority voting among its k nearest neighbors in the PCA-transformed feature space:

$$\hat{y} = \arg\max_{c \in \mathscr{C}} \sum_{i \in \mathcal{N}_k(\mathbf{x})} 1(y_i = c), \tag{21}$$

where $\mathbf{x}$ is the PCA-transformed feature vector of the sample to be classified, $\hat{y}$ is the predicted class label, $\mathscr{C}$ is the set of all possible classes, and $\mathcal{N}_k(\mathbf{x})$ represents the set of indices of the k nearest neighbors of $\mathbf{x}$ in the PCA-transformed space. The indicator function $\mathbf{1}(y_i = c)$ equals 1 if the i-th neighbor's class label y_i matches class c, and 0 otherwise.

Finally, Fig. 7D reports LR accuracy as a function of penalty type (L_1 or L_2) and regularization strength (C). Both penalty types achieve high accuracies, exceeding 0.96 for moderate values of C (0.1 to 1.0), demonstrating the effectiveness of these regularization strategies in mitigating overfitting in high-dimensional spectral data. Following hyperparameter tuning, the models were evaluated using various feature selection methods to assess their classification performance systematically.

Table 2, summarizes classification performance metrics for all four models and three feature selection methods, totaling twelve combinations. For each, results are reported separately for the test and training sets, along with the mean and standard

Table 2 Classification performance metrics for multiple models and feature selection methods[a]

Model	Feature selection method	Parameter	Accuracy	Precision	Recall	F1
RF	PCA	Testing	0.842 ± 0.041	0.891 ± 0.039	0.831 ± 0.051	0.834 ± 0.044
		Training	1.00	1.00	1.00	1.00
	RF-FS	Testing	0.886 ± 0.048	0.912 ± 0.039	0.867 ± 0.59	0.879 ± 0.050
		Training	1.00	1.00	1.00	1.00
	VT	Testing	0.886 ± 0.031	0.89 ± 0.031	0.86 ± 0.041	0.878 ± 0.032
		Training	1.00	1.00	1.00	1.00
SVM	PCA	Testing	0.854 ± 0.032	0.872 ± 0.025	0.837 ± 0.0573	0.849 ± 0.036
		Training	1.00	1.00	1.00	1.00
	RF-FS	Testing	0.893 ± 0.039	0.871 ± 0.023	0.861 ± 0.048	0.883 ± 0.046
		Training	1.00	1.00	1.00	1.00
	VT	Testing	0.833 ± 0.019	0.814 ± 0.032	0.804 ± 0.0405	0.819 ± 0.022
		Training	1.00	1.00	1.00	1.00
KNN	PCA	Testing	0.724 ± 0.053	0.762 ± 0.044	0.731 ± 0.040	0.707 ± 0.059
		Training	1.00	1.00	1.00	1.00
	RF-FS	Testing	0.862 ± 0.033	0.854 ± 0.0424	0.8323 ± 0.0479	0.851 ± 0.035
		Training	1.00	1.00	1.00	1.00
	VT	Testing	0.614 ± 0.060	0.72 ± 0.034	0.67 ± 0.052	0.607 ± 0.061
		Training	1.00	1.00	1.00	1.00
LR	PCA	Testing	0.850 ± 0.030	0.878 ± 0.016	0.846 ± 0.048	0.841 ± 0.032
		Training	1.00	1.00	1.00	1.00
	RF-FS	Testing	0.891 ± 0.046	0.91 ± 0.033	0.874 ± 0.060	0.884 ± 0.051
		Training	1.00	1.00	1.00	1.00
	VT	Testing	0.857 ± 0.025	0.841 ± 0.015	0.83 ± 0.036	0.847 ± 0.028
		Training	1.00	1.00	1.00	1.00

[a] Note: PCA: Principal Component Analysis; RF-FS: RF Feature Selection; VT: Variance Threshold.

deviation across the cross-validation folds. Among all methods, RF and LR combined with RF-based feature selection yielded the highest test accuracy, precision, recall, and F1 scores, all of which approached or exceeded 0.88. These results indicate that these models effectively classify mineral spectra, balancing false positives and false negatives well, and demonstrate robust generalization, as seen in the perfect training accuracy (reflecting model capacity) and slightly lower but reliable test performance. This pattern of near-perfect training accuracy coupled with slightly lower test accuracy, which reflects strong model capacity and reliable generalization, has also been observed in a spectroscopic classification study.[49]

SVM also performed well, especially with RF-FS (test accuracy 0.893 ± 0.039), although it was slightly lower than RF. SVM with PCA also maintained strong, balanced metrics, demonstrating the utility of PCA for dimensionality reduction. KNN showed lower test accuracy and F1 scores when using PCA or VT (reaching 0.724 ± 0.053 and 0.614 ± 0.060, respectively). Still, performance improved substantially with RF-FS (accuracy 0.862 ± 0.033), suggesting that KNN benefits significantly from supervised feature selection in this context. VT as a feature selector generally gave lower scores across all models compared to PCA or RF-FS. This illustrates the importance of utilising methods that either leverage label information, such as RF-FS, or preserve the overall variance structure, like PCA, for this type of data. Also, precision, recall, and F1 generally follow the same pattern as accuracy: when a model has higher accuracy, it usually means it's also good at minimizing both false positives and false negatives. This leads to high F1 scores and shows that the models aren't favoring any one class over others.

The results in Table 2 are supported by Fig. 8, which compares balanced accuracy across models and feature selection methods. Balanced accuracy is significant for this multi-class, imbalanced LIBS dataset, as it reflects the average recall across all mineral classes and prevents performance from being dominated by the largest class. As shown in Fig. 8, RF and LR combined with RF-based feature selection consistently achieve the highest and most stable balanced accuracy, with median values at or above 0.90 and very low variability, aligning with their strong performance in accuracy, precision, recall, and F1 scores reported in Table 2. SVM, especially when combined with RF-FS or PCA, also performs well, although it typically yields results below those of RF and LR. By contrast, KNN shows lower and more variable balanced accuracy with unsupervised selection (VT), but improves substantially when paired with RF-FS, highlighting the importance of label-informed feature selection. Models using VT alone tend to underperform, indicating that relying solely on global variance is insufficient for identifying informative spectral features. These findings underscore the value of supervised feature selection, especially RF-FS, for maximizing the generalizability and robustness of classification models in challenging, imbalanced mineral datasets.Fig. 9 displays the normalized confusion matrices for each classifier using RF-based feature selection, providing a granular view of class-level prediction performance. In these matrices, each row corresponds to the true mineral class, and each column to the predicted class. The values along the main diagonal (from top left to bottom right) represent the proportion of samples that were correctly classified for each mineral class; higher diagonal values indicate stronger model performance for those classes.

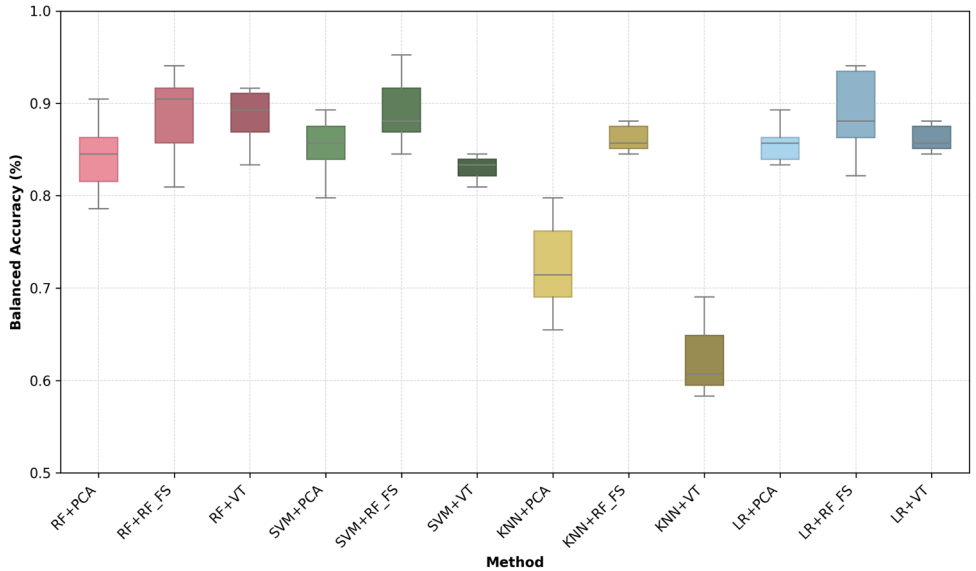


Fig. 8 Balanced accuracy distributions for each model and feature selection method.

Off-diagonal values, by contrast, indicate misclassifications where samples are incorrectly assigned to another class. Ideally, a perfect classifier would produce a matrix with all values on the diagonal and zeros elsewhere, while off-diagonal entries signal which minerals are most frequently confused. Fig. 9A confirms that RF achieves the most consistent and accurate mineral identification, with nearly all samples assigned to their correct class, reflected in minimal off-diagonal errors. Fig. 9B shows SVM, which retains a dominant diagonal but exhibits more class confusion than RF, especially for spectrally similar minerals, revealing specific class pairs that remain challenging to separate. Fig. 9C, for KNN, highlights more frequent misclassifications, particularly among classes with overlapping features, illustrating KNN's sensitivity to local variations and validating its comparatively lower balanced accuracy. Fig. 9D, LR, demonstrates performance close to RF, with most predictions along the diagonal and only occasional confusion between certain classes. Together, these confusion matrices not only validate the high overall accuracy of RF and LR with supervised feature selection but also pinpoint specific mineral classes where misclassification persists, providing actionable insight for refining future models and experimental design.

Fig. 10 shows the average Receiver Operating Characteristic (ROC) curves for each classifier and feature selection method, computed by averaging results over 10 random stratified splits. In these plots, the true positive rate (sensitivity) is plotted against the false positive rate (1-specificity) for varying classification thresholds. The proximity of the curve to the upper-left corner indicates stronger overall performance.

Each row of panels corresponds to a different classifier: Fig. 10A presents RF, Fig. 10B SVM, Fig. 10C KNN, and Fig. 10D LR. For RF, the ROC curves are consistently closest to the ideal point, with high area under the curve (AUC) values across all feature selection methods, reaffirming its robust discrimination ability observed in previous metrics and confusion matrices. SVM shows strong performance, particularly with RF-based feature selection, though with slightly more variability than that of RF. KNN exhibits noticeably flatter ROC curves, indicating weaker class separation and lower overall sensitivity, which aligns with its lower test accuracy and increased off-diagonal confusion. LR performs comparably to RF, especially when paired with supervised feature selection, demonstrating a high AUC and reliable classification boundaries.

These results provide consistent evidence of classifier performance across multiple evaluation criteria. Accuracy and F1 score, previously defined, measure overall correctness and balance between precision and recall, respectively.

The confusion matrix visually complements these metrics, with a strong diagonal indicating high true positive rates and minimal misclassifications. To further generalize classifier evaluation, ROC curves plot the true positive rate (TPR) against the false positive rate (FPR) across thresholds, where:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad \text{and} \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{22}$$

The area under the ROC curve (AUC) summarizes performance independent of threshold choice. As seen in Table 2, RF and LR with RF-based feature selection achieve the highest test accuracy, F1 score, and recall, as reflected in the confusion matrices, where nearly all predictions are on the diagonal, and the ROC curves approach an AUC of 1.0, indicating excellent discrimination. In contrast, models with lower accuracy and F1 scores, such as KNN with unsupervised feature selection, exhibit increased confusion and flatter ROC curves, indicating higher misclassification rates.

Overall, the agreement across accuracy, F1, confusion matrices, and ROC/AUC, grounded in their mathematical definitions, robustly validates the superior performance of RF and LR for multiclass LIBS mineral classification. The goal of supervised

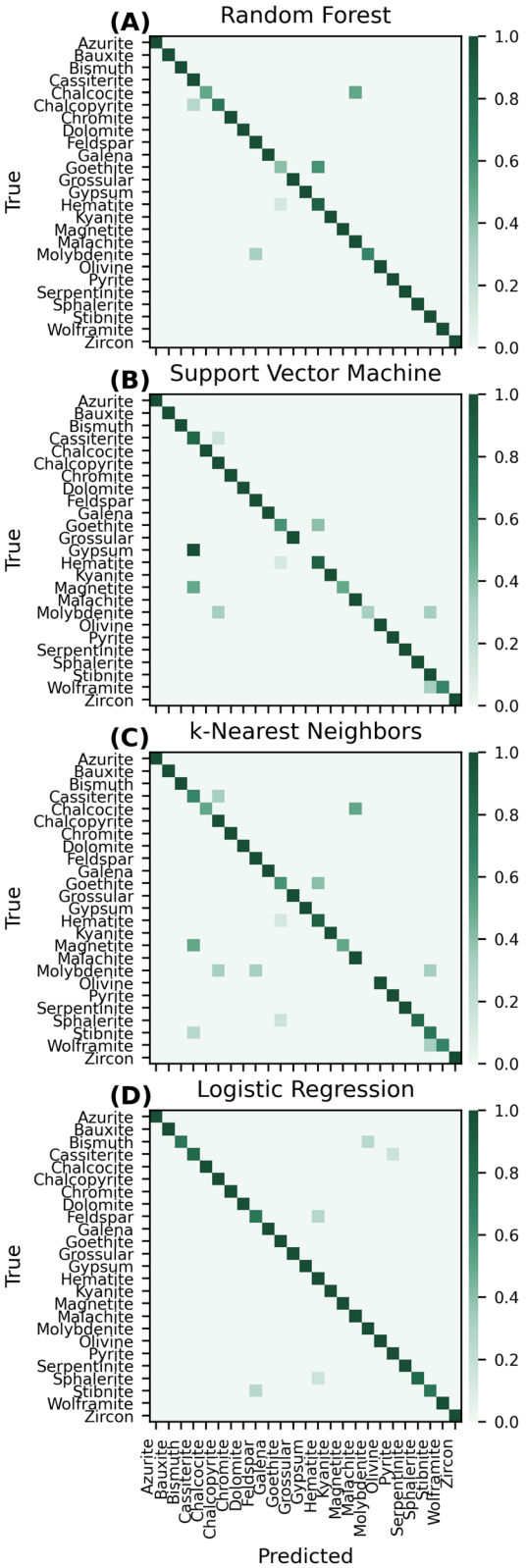


Fig. 9 Normalized confusion matrices for each classifier using RF-based feature selection: (A) RF, (B) SVM, (C) KNN, (D) LR. Strong diagonal values indicate high accuracy, while off-diagonal values highlight misclassifications between mineral classes.

pattern recognition is to use samples with known classes as a training set to build a model that can accurately predict the class of unknown samples. To achieve this, we first trained and validated our classifiers to achieve high performance on known data. To further evaluate their robustness and generalization, we then tested the best-performing models on 12 completely unseen spectra, which were randomly selected using a stratified sampling approach from the entire set of measured spectra. These new samples were processed with the same preprocessing steps as the training and testing data to maintain consistency. The models correctly classified 10 to 11 out of 12 samples, achieving an accuracy of approximately 83% to 92%, demonstrating strong predictive capability beyond the original dataset. Fig. 11 shows a comparison between the spectrum of an unseen bismuth sample and the mean spectrum of the predicted bismuth family, as classified by the LR model with RF-based feature selection. The close alignment of key spectral peaks between the individual sample and the family mean spectrum highlights the model's ability to generalize and classify new, unseen spectra accurately. This visual confirmation supports the quantitative classification results, demonstrating the robustness of this approach for real-world mineral identification.This strong performance on previously unseen pure mineral samples demonstrates the practical potential of our approach for autonomous mineral identification. However, planetary materials and terrestrial soils are rarely pure phases and are often complex mixtures of several minerals. Therefore, to further test the robustness and interpretability of our models under realistic conditions, we systematically evaluated classifier performance on synthetic binary mixtures, as described below.

3.1 Classification of synthetic mineral mixtures

To address the robustness of our classifier on challenging, real-world samples that are mixtures of minerals (as present in soils and rocks), we generated synthetic mixture spectra by linearly combining measured spectra of hematite and gypsum in varying proportions. Specifically, for two minerals with normalized spectra $S_1(\lambda)$ and $S_2(\lambda)$, a mixture spectrum was computed as follows:

$$S_{\text{mix}}(\lambda) = w_1 S_1(\lambda) + w_2 S_2(\lambda) \quad (23)$$

where w_1 and $w_2 = 1 - w_1$ denote the fractions of hematite and gypsum, respectively. This procedure was repeated for w_1 from 0.1 to 0.9 in increments of 0.1.

Synthetic mixture spectra, together with pure hematite and gypsum spectra measured at 10 mbar, are shown in Fig. 12. Progressive changes in spectral features with varying composition indicate compositional sensitivity and experimental relevance of the mixtures. For the mixture classification analysis, we report results using the LR-FS model, identified as one of the best-performing models in this study (see Table 2). Each synthetic mixture was classified with this model, and the probability assigned to hematite, $P_{(\text{Hematite})}$, was recorded. To quantitatively analyze the classifier's response to these mixtures, the following metrics were evaluated.

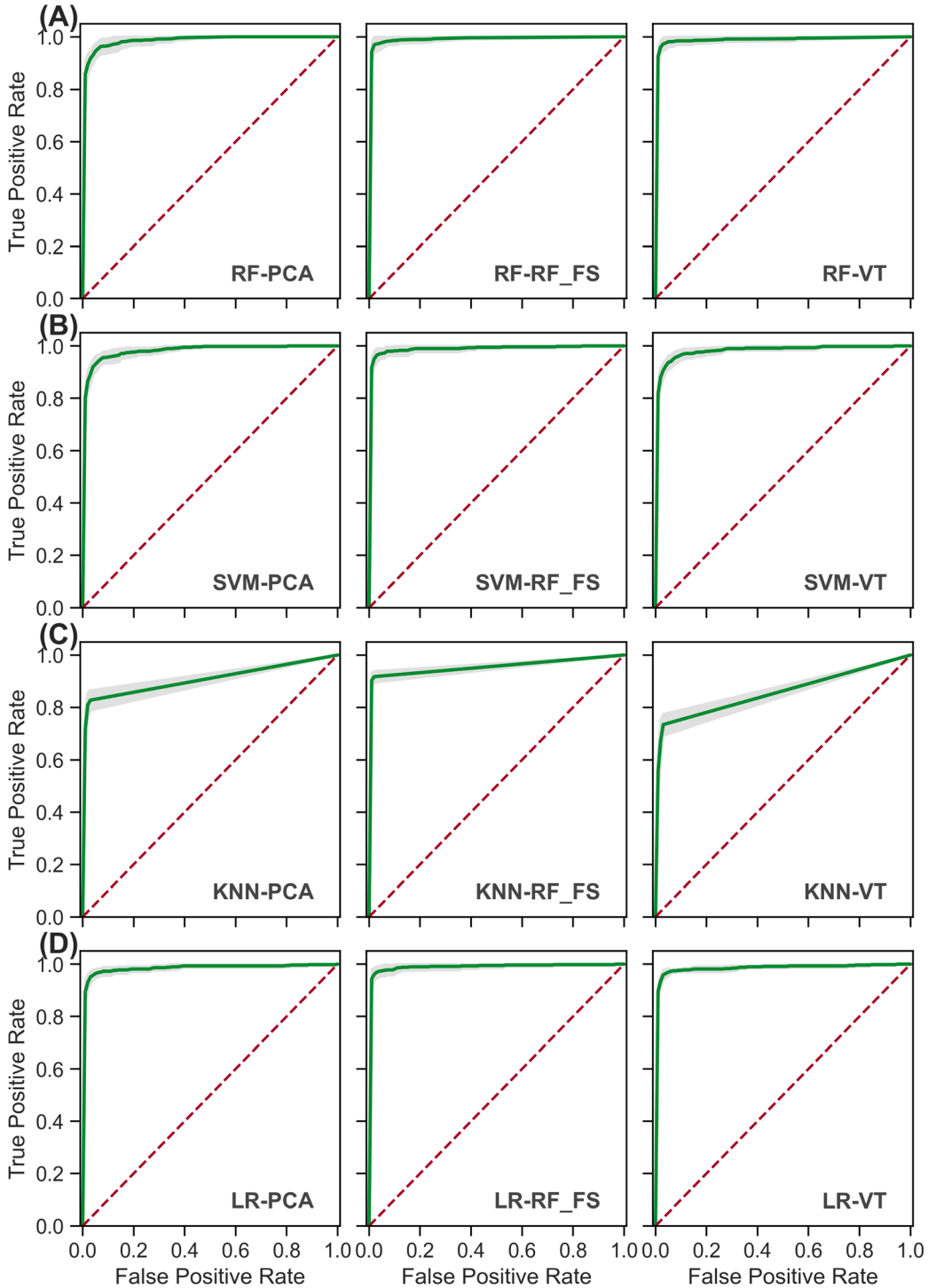


Fig. 10 Average ROC curves for each classifier (A) RF, (B) SVM, (C) KNN, and (D) LR, and feature selection method, computed across 10 random splits. Green lines show the mean ROC, grey bands represent ± 1 standard deviation, and the red dashed line indicates the chance level. Curves closer to the top-left demonstrate stronger classification performance.

First, the relationship between the predicted probability P_{Hematite} and its fraction in the mixture f_{Hematite} was described using a sigmoid function:

$$P_{\text{Hematite}}(f_{\text{Hematite}}) = \alpha + (1-\alpha)\cdot\frac{1}{1+\exp(-k\cdot(f_{\text{Hematite}}-f_0))} \quad (24)$$

where f_0 is the inflection ("switch") point indicating the mixture ratio at which the predicted class transitions, k controls the steepness of this transition, and α allows for a nonzero baseline probability.

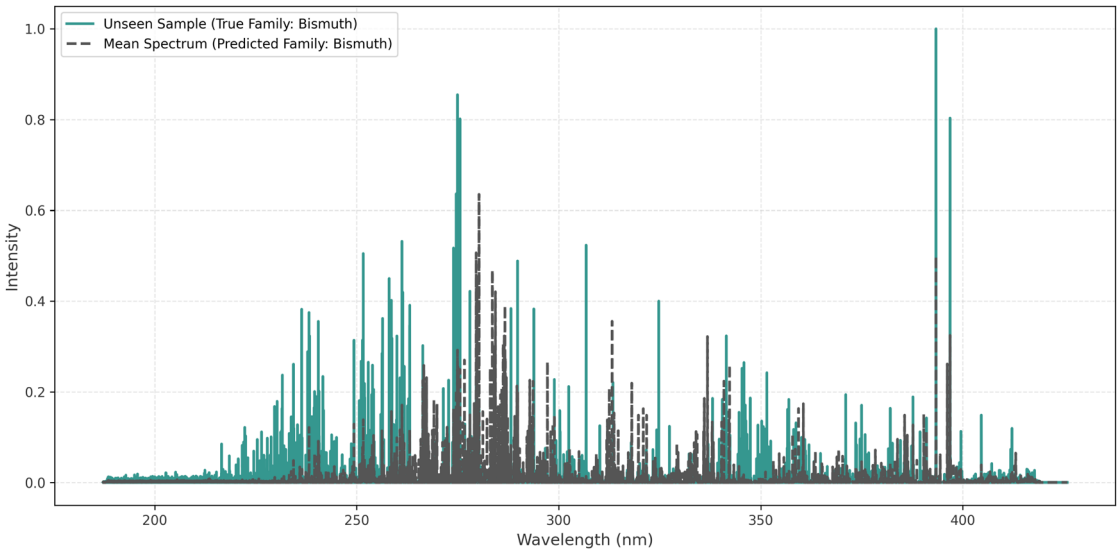


Fig. 11 Comparison of an unseen bismuth sample spectrum (green) with the mean predicted family spectrum (grey) classified by LR + RF-FS, shown here as an example of model performance on new data.

Then, to evaluate prediction confidence, the entropy H of the predicted probability distribution across all mineral classes was computed for each mixture:

$$\mathrm{H} = -\sum_i P_i \log P_i \quad (25)$$

where P_i is the assigned probability for class i. Higher entropy values show greater uncertainty in the classification.

Finally, the ability to identify the dominant mineral in each mixture was assessed by constructing an ROC curve, using the probability assigned to hematite (P_{Hematite}) as the score and the dominant mineral as ground truth. The AUC provides a summary measure of discriminative performance for mixed samples.

Fig. 13 shows how our best classifier (LR + RF-FS) performs on synthetic mixtures of hematite and gypsum, with clear evidence of both accuracy and reliability in mixed-mineral

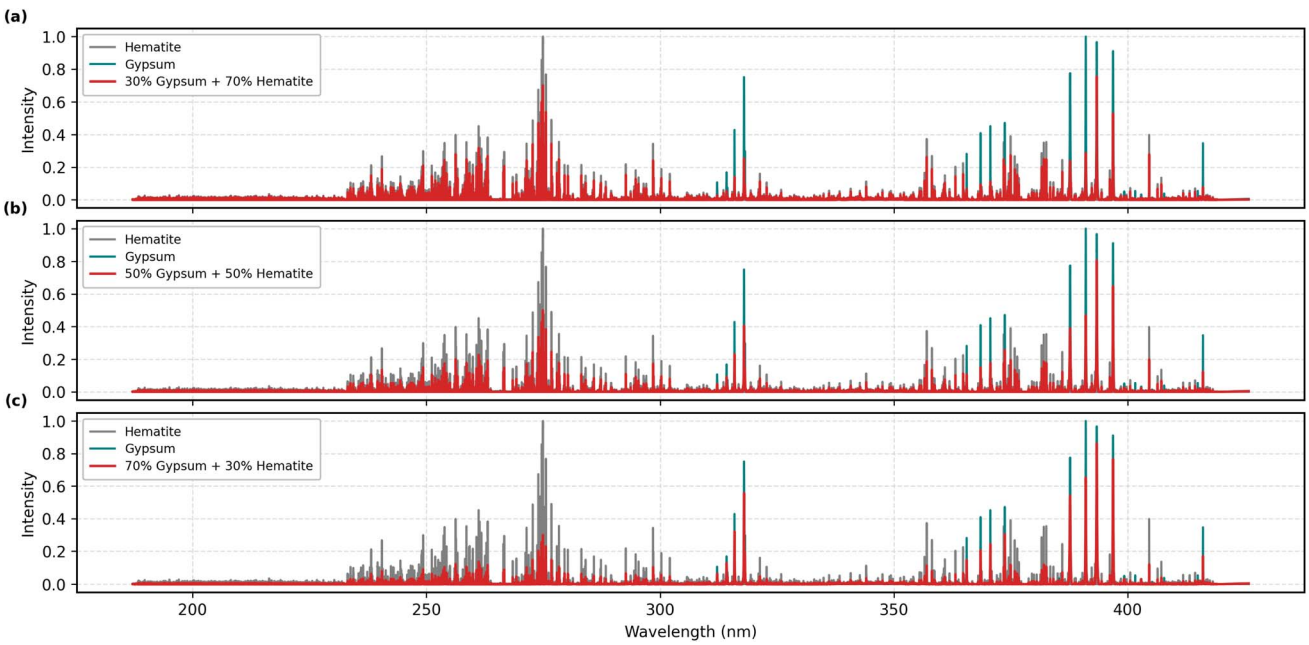


Fig. 12 Example of synthetic mixture spectra generated by linearly combining normalized spectra of hematite and gypsum, all measured at a pressure of 10 mbar. (A) 70% hematite + 30% gypsum, (B) 50% hematite + 50% gypsum, and (C) 30% hematite + 70% gypsum. Pure hematite and gypsum spectra at 10 mbar are also shown for reference in each panel. The gradual transition of spectral features demonstrates the compositional sensitivity of the mixture and validates the approach for simulating realistic mixed mineral samples under controlled pressure conditions.

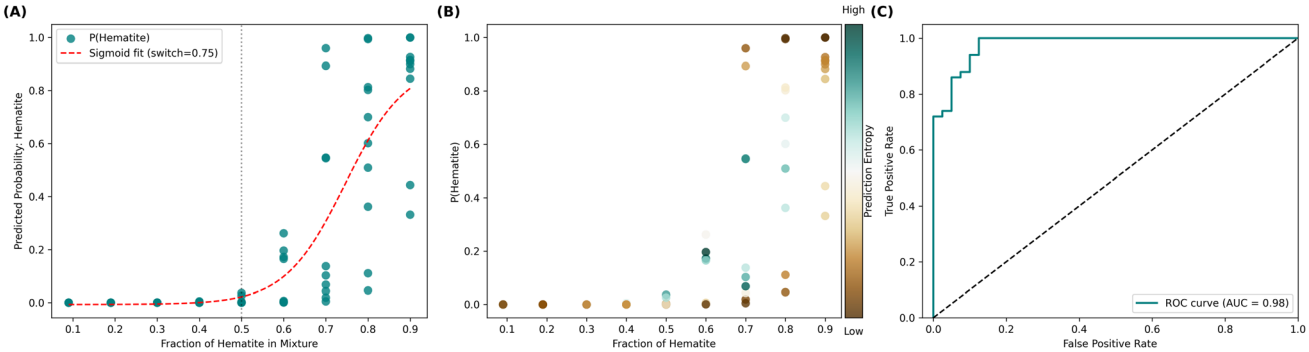


Fig. 13 Classifier performance and prediction confidence on synthetic hematite–gypsum mixtures. (A) Predicted probability for hematite as a function of its fraction in the mixture, with a sigmoid fit (red dashed line). (B) Probability *vs.* hematite fraction, colored by prediction entropy. (C) ROC curve for dominant mineral identification (AUC = 0.98). Results are shown for the LR + RF-FS classifier.

samples. Panel A shows the relationship between the predicted probability for hematite and its actual fraction in the mixture. Each dot is a single synthetic mixture (created by mixing hematite and gypsum spectra in different amounts), and the red dashed line is a sigmoid curve fitted to the data. The "switch point" of the curve is at about 0.75, meaning that the model begins to call the mixture "hematite" when hematite is roughly 75% of the sample. This curve confirms that the model responds in a logical and gradual way to changing mineral composition, rather than making random or abrupt jumps. Panel B shows the same probability values, but now the color of each point represents the prediction entropy, which measures the classifier's confidence. When the mixture is nearly all hematite or all gypsum (far left or right on the *x*-axis), the classifier is very confident (entropy is low, brown/yellow points). The highest uncertainty (blue points) is found near the middle, where hematite and gypsum are in similar amounts, making the classification harder. Panel C displays the ROC curve, which evaluates how well the classifier can identify the dominant mineral in each mixture using the predicted probability for hematite. The curve is very close to the top left corner, and the AUC is 0.98, indicating overall high accuracy, particularly when one mineral dominates. However, as seen in Fig. 13A, the classifier's predictions are less reliable for intermediate mixtures where the proportions of hematite and gypsum are similar. The model also provides interpretable confidence estimates. This capability is especially important for real-world applications in soils, rocks, and planetary materials, where mixtures are common and confident decisions are required.

However, true deployment in planetary missions brings further challenges, such as the need to recognize minerals not present in the training library and to support autonomous, onboard decision-making. These aspects are addressed in the following section.

3.2 Addressing unknown spectra and onboard decision-making

To identify minerals from LIBS data, machine learning models typically require supervised training using extensive libraries of known mineral spectra. However, when exploring other planets, scientists often encounter minerals that are not included in these existing datasets. To address this, our method employs a learning approach that enables the model to recognise unfamiliar spectra and suggest potential matches. At the same time, it can update and grow the spectral library by adding new data as it becomes available. This highlights the importance of continually expanding and diversifying spectral databases, particularly for space mining. Another significant advantage arises from the establishment of a feedback loop between the LIBS instrument and the classification system. If the model is not confident in its classification or cannot assign a spectrum, the system can automatically trigger additional scans or measurements. This back-and-forth process helps improve accuracy and allows faster, more reliable decisions to be made directly on the spacecraft. Such real-time feedback is especially crucial in space missions, where communication delays with Earth prevent immediate human input. Furthermore, robust and understandable mineral classification is made possible by the interpretability offered by random forest-based feature selection (RF-FS), which identifies particular elemental emission lines rather than abstract features. Onboard decision-making in extraterrestrial environments requires autonomous systems to handle unseen mineral spectra, dynamically adjust measurement strategies, and make accurate, real-time resource assessments. Finally, although computing power on a spacecraft is more limited than on Earth, current models are designed to be compact and efficient enough to run on small or embedded computers onboard. Thanks to improvements in lightweight computing hardware, it is possible to quickly identify minerals and assess resources right there on the spacecraft. This capability is vital to enable autonomous decisions during future robotic and crewed missions. However, there are challenges ahead in handling large spectral libraries and updating models during missions, which will be important areas of ongoing work.

4 Conclusion

This study evaluated multiple ML models combined with various feature selection methods for LIBS mineral

classification. The RF and LR models combined with RF-FS achieved the highest test accuracies of approximately 88.6% and 89.1%, respectively. SVM with RF-FS also performed well, with a test accuracy around 89.3%. KNN demonstrated moderate performance, achieving up to 86.2% accuracy when combined with RF-FS, but generally yielded lower results with unsupervised feature selectors. Models using PCA and VT for feature selection yielded slightly lower accuracies overall. Class imbalance was addressed using SMOTE, and hyperparameter tuning was performed by a grid search to optimize model parameters and enhance classification performance. The fundamental concept of supervised pattern recognition is to use samples with known classes as a training set to build a model that can predict the class of unknown samples. Building on this principle, to further evaluate model generalization beyond the cross-validation framework, the best-performing classifiers were tested on 12 completely unseen spectra, randomly selected using a stratified sampling approach from the full dataset. These new spectra were subjected to the identical preprocessing pipeline to ensure consistency. The models correctly classified 10 to 11 out of the 12 samples, corresponding to an accuracy of approximately 83% to 92%. This shows robust predictive capacity and supports the practical applicability of the LIBS-ML approach to mineral identification in real-world scenarios, such as autonomous space mining. The approach also offers reliable mixture classification, interpretability through emission line matching, and confidence estimation for autonomous adaptation to new spectra.

Author contributions

HS: conceptualisation, formal analysis, writing – original draft preparation, investigation, supervision, visualisation, writing – review & editing. PŠ: data curation. AK. M, JŽ, BA, PŠ, and PK: writing – review & editing. MF: funding acquisition, project administration, resources, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5ja00377f.

Acknowledgements

HS and PK would like to acknowledge the financial support of the Czech Science Foundation (GACR) under grant number 23-05186 K. MF and HS would like to acknowledge the financial support of the Technology Agency of the Czech Republic (TAČR) under the grant of the National Centre of Competence for Aeronautics and Space, reg. No. TN02000009/11 FREYA II. B.A. and J.Z. are grateful for funding *via* the EU Project ERA Chair – SPACE – Heyrovsky Chair of Space Science, no. 101186661.

Notes and references

1 D. Comelli, M. D'orazio, L. Folco, M. El-Halwagy, T. Frizzi, R. Alberti, V. Capogrosso, A. Elnaggar, H. Hassan, A. Nevin, F. Porcelli, M. G. Rashed and G. Valentini, *Meteorit. Planet. Sci.*, 2016, **51**, 1301–1309.
2 R. Verduci, V. Romano, G. Brunetti, N. Yaghoobi Nia, A. Di Carlo, G. D'Angelo and C. Ciminelli, *Adv. Energy Mater.*, 2022, **12**, 2200125.
3 K. Tsiolkovsky, *Scientific Review*, Moscow, St. Petersburg, 1903, 221, p. 222.
4 A. C. Clarke, *J. Br. Interplanet. Soc.*, 1950, **9**, 261–267.
5 J. Cilliers, K. Hadler and J. Rasera, *npj Microgravity*, 2023, **9**, 22.
6 C. Whetsel, J. S. Levine, S. J. Hoffman, C. M. Luckey, K. D. Watts and E. L. Antonsen, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**, e2404251121.
7 C. D. Espejel, S. Casanova, S. Saydam and J.-A. Lamamy, in *Handbook of Space Resources*, ed. V. Badescu, K. Zacny and Y. Bar-Cohen, Springer International Publishing, 2023, pp. 999–1022.
8 G. F. Sowers and C. B. Dreyer, *New Space*, 2019, **7**, 235–244.
9 A. Meurisse and J. Carpenter, *Planet. Space Sci.*, 2020, **182**, 104853.
10 K. M. Cannon and D. T. Britt, *Icarus*, 2020, **347**, 113778.
11 J. A. Hoffman, M. H. Hecht, D. Rapp, J. J. Hartvigsen, J. G. SooHoo, A. M. Aboobaker, J. B. McClean, A. M. Liu, E. D. Hinterman, M. Nasr, *et al.*, *Sci. Adv.*, 2022, **8**, eabp8636.
12 H. Saeidfirozeh, P. Kubelík, V. Laitl, A. Křivková, J. Vrábel, K. Rammelkamp, S. Schröder, I. Gornushkin, E. Képeš, J. Žabka, *et al.*, *TrAC, Trends Anal. Chem.*, 2024, 117991.
13 S. Maurice, R. Wiens, M. Saccoccio, B. Barraclough, O. Gasnault, O. Forni, N. Mangold, D. Baratoux, S. Bender, G. Berger, *et al.*, *Space Sci. Rev.*, 2012, **170**, 95–166.
14 W. Xu, X. Liu, Z. Yan, L. Li, Z. Zhang, Y. Kuang, H. Jiang, H. Yu, F. Yang, C. Liu, *et al.*, *Space Sci. Rev.*, 2021, **217**, 64.
15 A. Laxmiprasad, R. Sridhar, A. Goswami, K. Lohar, M. Rao, K. Shila, M. Mahajan, B. Raha, T. Smaran and B. Krishnamprasad, *Curr. Sci.*, 2020, **118**, 573–581.
16 H. Y. McSween Jr and M. H. Thiemens, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**, e2415280121.
17 E. Képeš, H. Saeidfirozeh, V. Laitl, J. Vrábel, P. Kubelik, P. Pořízka, M. Ferus and J. Kaiser, *J. Anal. At. Spectrom.*, 2024, **39**, 1160–1174.
18 H. Saeidfirozeh, A. K. Myakalwar, P. Kubelík, A. Ghaderi, V. Laitl, L. Petera, P. B. Rimmer, O. Shorttle, A. N. Heays, A. Křivková, *et al.*, *J. Anal. At. Spectrom.*, 2022, **37**, 1815–1823.
19 V. Payré, C. Fabre, V. Sautter, A. Cousin, N. Mangold, L. Le Deit, O. Forni, W. Goetz, R. C. Wiens, O. Gasnault, *et al.*, *Icarus*, 2019, **321**, 736–751.
20 R. J. Baumgartner, M. L. Fiorentini, D. Baratoux, L. Ferrière, M. Locmelis, A. Tomkins and K. A. Sener, *Meteorit. Planet. Sci.*, 2017, **52**, 333–350.

21 E. Chatzitheodoridis, P. Clerc, A. Kereszturi, N. Mason, E. Persson, C. Possnig, L. Poulet, M. Puumala, O. Sivula and J. R. Brucato *et al.*, in *Mars and the Earthlings: A Realistic View on Mars Exploration and Settlement*, Springer, 2025, pp. 253–339.
22 M. Crater, *science*, 2014, 343.
23 C. Popa, G. Carrozzo, G. DiAchille, S. Silvestro, F. Espostio and V. Mennella, *EGU General Assembly Conference Abstracts*, 201512056.
24 C. Gil-Lozano, E. Mateo-Marti, L. Gago-Duport, E. Losa-Adams, M. F. Sampedro, J. Bishop, V. Chevrier and A. G. Fairén, *Front. Astron. Space Sci.*, 2025, **12**, 1504288.
25 J. Gillespie, A. J. Cavosie, D. Fougerouse, C. L. Ciobanu, W. D. Rickard, D. W. Saxey, G. K. Benedix and P. A. Bland, *Sci. Adv.*, 2024, **10**, eadq3694.
26 B. L. Ehlmann, J. F. Mustard, S. L. Murchie, F. Poulet, J. L. Bishop, A. J. Brown, W. M. Calvin, R. N. Clark, D. J. D. Marais, R. E. Milliken, *et al.*, *Science*, 2008, **322**, 1828–1832.
27 H. King, Gypsum Mineral|Uses and Properties, Geology.com, 2019, https://geology.com/minerals/gypsum.shtml, retrieved 31 March 2019.
28 P. B. Niles, D. C. Catling, G. Berger, E. Chassefière, B. L. Ehlmann, J. R. Michalski, R. Morris, S. W. Ruff and B. Sutter, *Space Sci. Rev.*, 2013, **174**, 301–328.
29 Minerals.cz, https://www.minerals.cz.
30 D. W. Hahn and N. Omenetto, *Appl. Spectrosc.*, 2012, **66**, 347–419.
31 P. Pořízka, J. Klus, E. Képeš, D. Prochazka, D. W. Hahn and J. Kaiser, *Spectrochim. Acta, Part B*, 2018, **148**, 65–82.
32 B. Zhang, L. Sun, H. Yu, Y. Xin and Z. Cong, *J. Anal. At. Spectrom.*, 2013, **28**, 1884–1893.
33 Y. Sun, L. Liu and L. Xiao, *3rd International Conference on Electronics and Information Technology (EIT)*, 2024, pp. 175–178.
34 G. R. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt and A. Leary, *J. Open Source Softw.*, 2019, **4**, 1237.
35 M. Kuhn and K. Johnson *et al.*, *Applied Predictive Modeling*, Springer, 2013, vol. 26.
36 N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, *J. Artif. Intell. Res.*, 2002, **16**, 321–357.
37 S. Arlot and A. Celisse, *Stat. Surv.*, 2010, **4**, 40–79.
38 A. Kramida, Y. Ralchenko and J. Reader, *NIST ASD Team, NIST Atomic Spectra Database*, version 5.9, National Institute of Standards and Technology, 2021, Gaithersburg, MD, https://physics.nist.gov/asd.
39 I. Karnadi, M. Pardede, E. Harefa, I. Tanra, R. Hedwig, B. Harsono, M. Y. Hadiyanto, T. T. Lie, W. Zhou, K. Kagawa, *et al.*, *Opt. Continuum*, 2023, **2**, 1028–1039.
40 J. Sumathi, V. S. Kumar and K. Veerappan, *J. Appl. Spectrosc.*, 2022, **88**, 1215–1228.
41 Z. Hao, K. Liu, Q. Lian, W. Song, Z. Hou, R. Zhang, Q. Wang, C. Sun, X. Li and Z. Wang, *Front. Phys.*, 2024, **19**, 62501.
42 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
43 V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.
44 T. Cover and P. Hart, *IEEE Trans. Inf. Theory*, 1967, **13**, 21–27.
45 D. R. Cox, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 1958, **20**, 215–232.
46 J. Gorodkin, *Comput. Biol. Chem.*, 2004, **28**, 367–374.
47 F. J. Massey Jr, *J. Am. Stat. Assoc.*, 1951, **46**, 68–78.
48 L. Sheng, T. Zhang, G. Niu, K. Wang, H. Tang, Y. Duan and H. Li, *J. Anal. At. Spectrom.*, 2015, **30**, 453–458.
49 C. Y. Tachie, D. Obiri-Ananey, M. Alfaro-Cordoba, N. A. Tawiah and A. N. Aryee, *Food Chem.*, 2024, **431**, 137077.