International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Exploring Spectrometric Biosignatures through Data Science

Prof. Gautam Dematti [a], ***Avantika Patil*** [b], ***Bhagyalaxmi Patil*** [c], ***Mayuri Patil*** [d], ***Sakshi barbari*** [e]

[a,b,c,d,e] Department of Artificial Intelligence and data Science, Angadi Institute of Technology and Management, Belagavi-590009, India

ABSTRACT

This project sits at the crossroads of data science and astrobiology, using advanced algorithms like Keras CNN and Tensorflow to analyze extensive datasets from Martian missions. Keras CNN identifies temporal patterns, providing insight into Martian processes, while Random Forest enhances the detection of subtle life-related indicators. Beyond Mars exploration, the research contributes to discussions on the habitability of celestial bodies, employing a systematic, data-driven approach to the search for extraterrestrial life. The interdisciplinary effort establishes a framework for future missions and analyses, aiming to advance our understanding of the Martian environment and humanity's quest for knowledge beyond Earth. The outcomes extend to broader discussions on life beyond our planet, shaping our perspective on our place in the cosmos.

Keywords: Keras CNN, Tensorflow, Data-driven approach, Extraterrestrial life etc.

Introduction

The project, "Detection of Life on Mars Using Data Science," represents a collaborative and interdisciplinary endeavor aimed at unraveling the mysteries of the Red Planet through the application of advanced data science techniques. The intricate task involves meticulously analyzing extensive datasets collected from Martian surfaces with a primary focus on identifying potential indicators of life.

At the core of our methodology is a commitment to a holistic approach, recognizing that the search for extraterrestrial life requires a comprehensive understanding of both the Martian environment and the intricate nuances of data analytics. The collaboration between different disciplines ensures that the project is not only technologically advanced but also grounded in the expertise necessary to interpret the findings accurately.

As we navigate the vast landscape of Martian data, one key aspect of our strategy is the careful consideration of temporal dependencies in time-series data. This approach enables us to discern nuanced patterns that may be indicative of potential biological or chemical activity. The temporal dimension is crucial in understanding the dynamic processes that could hint at the presence of life. By strategically employing methods to capture these temporal dependencies, we enhance our ability to interpret the data accurately and make informed conclusions.

Simultaneously, our project addresses the challenge of distinguishing between ordinary geological features and potential life-related processes. The Martian landscape is rich with geological complexities, and separating natural phenomena from potential signs of life requires a sophisticated analytical approach. While the specific algorithms employed remain implicit in this overview, they are designed to handle the intricacies of the Martian datasets with precision.

It's important to note that our project goes beyond the confines of algorithmic intricacies. We recognize the broader implications of our work in the context of advancing our understanding of the possibility of life on Mars. The exploration of extraterrestrial life is not solely a technological pursuit but also a

venture that prompts profound questions about our place in the universe and the potential for life beyond Earth.

By pushing the boundaries of what data science can achieve in the context of Martian exploration, our project contributes to a growing body of knowledge that extends beyond the immediate goal of detecting life on Mars. The advancements in data analytics and the insights gained from this endeavor have far-reaching implications for the broader field of astrobiology and our understanding of habitability within our solar system and beyond.

The collaborative nature of our interdisciplinary approach is a key strength. It not only reflects the complexity of the challenge at hand but also establishes a framework for future missions and analyses. The fusion of expertise from different fields not only enhances the scientific rigor of our work but also fosters a culture of collaboration and innovation that is essential for tackling the profound questions surrounding the existence of life beyond Earth.

Methodology

The Methodology for comparing CNN-frameworks for medicinal plant identification will depend on the specific CNN-frameworks being compared and the characteristics of the data. Here are a few general steps that might be involved in this process.

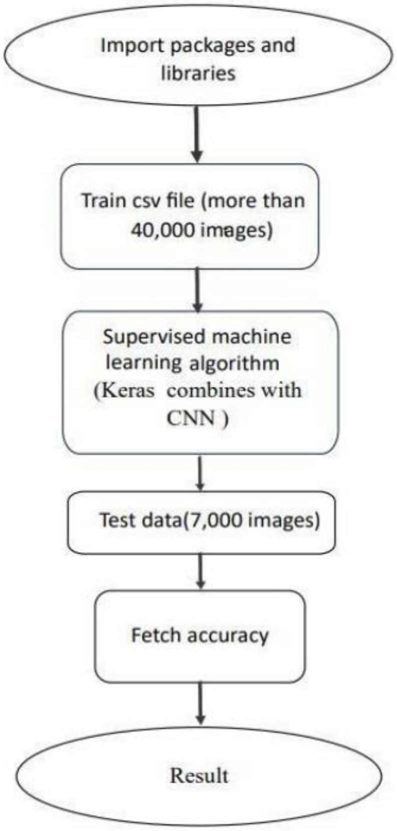

Import packages and libraries
Train csv file (more than 40,000 images)
Supervised machine learning algorithm (Keras combines with CNN)
Test data(7,000 images)
Fetch accuracy
Result

Fig. Block Schematic Diagram

Image acquisition: means taking pictures or capturing visual data using cameras, scanners or sensors. These devices transform real things or scenes into digital images that computers can store, process or understand. There are several ways to do this: cameras use lenses to convert light into digital images, scanners convert physical things into digital ones, sensors detect things like infrared or ultraviolet light to make digital images, and even our phones and tablets have cameras. pictures This whole process is very important in fields such as medicine, monitoring, photography and other fields. How clear, detailed and accurate the images are is really important because it affects how well computers work with them later.

Preprocessing: means preparing things before working with them. Images or information are cleaned or organized to make them easier to use. This may include, for example, removing errors or unwanted parts, changing sizes or formats to prepare the data for analysis or further processing. It's like cleaning before you start working to make everything smoother and more understandable.

Feature extraction: At this point, several functions are extracted from the file segmented image and effective features are selected classification of additional crops. Extraction of different features techniques give different results and detection Improving proper functions is one of the critical tasks of the factory identification.

Calculating the leaf factor: Involves working out the details of specific pages. It is like measuring or studying different aspects of leaves such as size, shape or texture. This information helps in understanding and classifying leaves, especially identifying different plants based on their unique leaf characteristics. These leaf counts or measurements are important in studies related to plant identification or classification.

Leaf Factor in Ayurvedic Database: Focuses on understanding the properties of leaves. It examines characteristics such as shape, size, structure and other details characteristic of medicinal plant leaves. This information helps identify and classify plants based on their leaf properties, supports the classification and study of plants used in Ayurvedic medicine based on their medicinal properties.

Sample images used in dataset:

This dataset contains 40,000 images taken by the Mars Science Laboratory (MSL) rover using three instruments: Mastcam Right eye, Mastcam Left eye, and MAHLI. The images are in a "browse" version, each approximately 256x256 pixels. For full-resolution images, access can be obtained from the PDS.

To facilitate operational use of the image archive over time, the dataset has been segregated into training, validation, and test sets based on the sol (Martian day) of acquisition. This division covers a range from sols 3 to 1060, and the specific breakdowns for the training, validation, and test sets are outlined in individual files.

Fig. 1: Sample figure 1

Fig. 2: Sample figure 2

Fig. 3: Sample figure 3

Fig. 4: Sample figure 4

System Design

Data Flow Diagram:

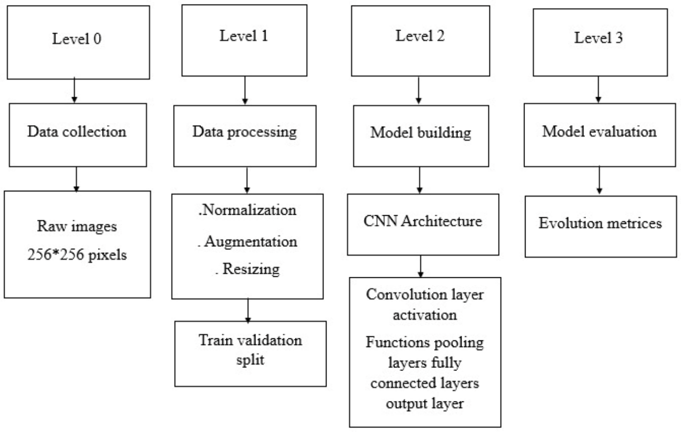


Fig.: Data flow Diagram

Level 0: Data Collection - This is the initial stage of the project, where raw images of Mars are collected. In this case, these images would be the dataset of Mars pictures you mentioned.

Level 1: Data Processing - Once the data is collected, it needs to be prepared for analysis. This stage involves processing the raw images, which could include resizing them to a standard size (e.g., 256x256 pixels) and normalizing the pixel values. Normalization is a common preprocessing step in machine learning that scales numerical data to a range that is more suitable for the algorithms to work with.

Level 2: Model Building - After preparing the data, the next step is to build a model for analyzing the data. In this case, the model is a Convolutional Neural Network (CNN), which is a type of deep learning architecture commonly used for image analysis tasks. The CNN architecture consists of various layers, including convolution layers (with activation functions), pooling layers, and fully connected layers. These layers work together to automatically learn and extract features from the input images, which can then be used to make predictions

Level 3: Model Evaluation - Once the model is built, it needs to be evaluated to ensure that it's performing well and can accurately analyse the Mars pictures. This stage involves splitting the dataset into training and validation sets, and then assessing the model's performance on the validation set. Various evaluation metrics, such as those related to evolution (e.g., fitness or accuracy), can be used to gauge the model's effectiveness.

Sequence diagram:

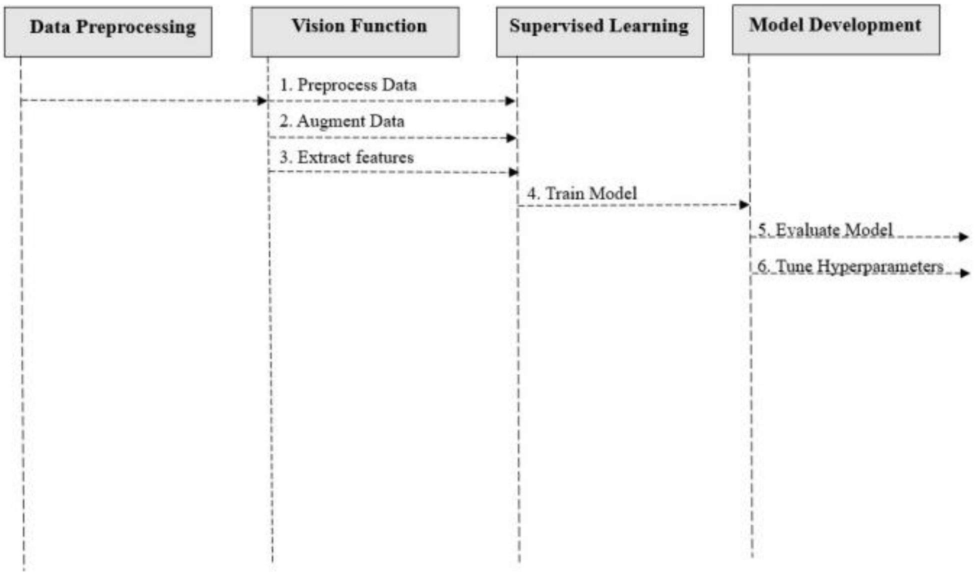


Fig. : Sequence Diagram

Preprocess Data: The first step in the machine learning pipeline involves preparing the raw data for analysis. For your project, this might mean filtering out irrelevant images, resizing images to a uniform size, and normalizing pixel values.

Augment Data: Data augmentation artificially expands your dataset by applying various transformations to existing images. This could include rotating, flipping, or zooming in and out on the images. These variations help make the machine learning model more robust and capable of generalizing better to new data.

Extract Features: Next, we extract relevant features from the preprocessed and augmented images. For image data, this typically involves using a pre-trained convolutional neural network (CNN) to pull out high-level features. These extracted features will serve as inputs for the machine learning model.

Train Model: With the features ready, we move on to training the machine learning model. For your project, this might mean training a binary classifier to differentiate between images that show signs of life and those that do not.

Evaluate Model: After training, the model is tested on a separate dataset to evaluate its performance. This step helps identify any issues with overfitting or underfitting and gives an estimate of how well the model will perform on new, unseen data.

Tune Hyperparameters: Finally, we optimize the model's performance by tuning its hyperparameters. This could involve adjusting settings like the learning rate or regularization strength to find the best configuration for your specific dataset.

Use Case Diagram :

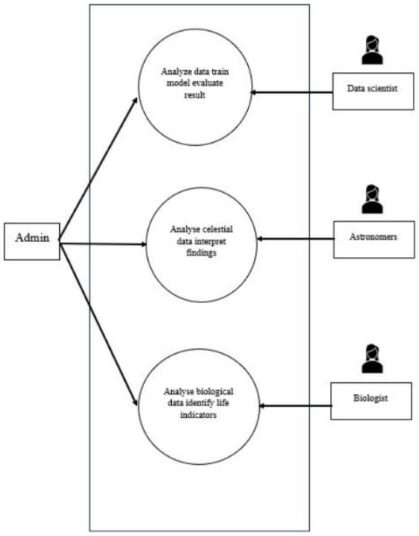

Analyze data train model evaluate result
Data scientist
Admin
Analyse celestial data interpret findings
Astronomers
Analyse biological data identify life indicators
Biologist

Fig. : Use Case Diagram

Admin: This use case likely represents the administrative tasks required to manage the project, such as user management, access control, and configuration settings.

Analyse celestial data interpret findings: This use case represents the process of analysing the celestial data obtained from the Mars pictures and interpreting the findings. Astronomers or data scientists may be responsible for this task.

Analyse biological data identify life indicators: This use case represents the process of analysing the biological data obtained from the Mars pictures to identify any indicators of life. Biologists or data scientists may be responsible for this task.

Data scientist, Astronomers, Biologist: These are the roles or personas involved in the project. Data scientists would be responsible for training and evaluating the machine learning model, astronomers would analyse the celestial data, and biologists would analyse the biological data.

Implementation:

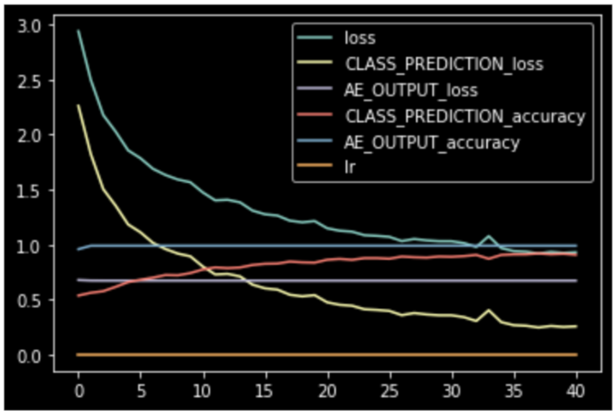
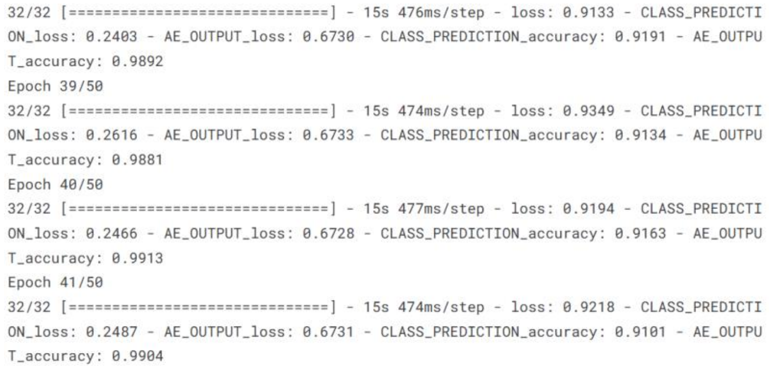


Fig. : Accuracy and Loss prediction

His graph appears to show the training progress of a neural network model over epochs. It includes the following metrics:

i. Loss (General Loss) - This measures the overall performance of the model, and it typically decreases as the model learns.
ii. CLASS_PREDICTION_loss - The loss specifically associated with the class prediction task, likely part of a multi-task learning model. This also tends to decrease as the model improves.
iii. AE_OUTPUT_loss - The loss for the autoencoder output, indicating how well the autoencoder part of the model is reconstructing the input. This should decrease with training.
iv. CLASS_PREDICTION_accuracy - The accuracy of the class predictions. This usually increases as the model learns.
v. AE_OUTPUT_accuracy - The accuracy of the autoencoder output, which could be interpreted in various ways depending on the specific task. It generally increases as the model improves.
vi. Lr (Learning Rate) - This shows the learning rate over epochs. It can be adjusted dynamically during training to optimize learning.



```
32/32 [==============================] - 15s 476ms/step - loss: 0.9133 - CLASS_PREDICTI
ON_loss: 0.2403 - AE_OUTPUT_loss: 0.6730 - CLASS_PREDICTION_accuracy: 0.9191 - AE_OUTPU
T_accuracy: 0.9892
Epoch 39/50
32/32 [==============================] - 15s 474ms/step - loss: 0.9349 - CLASS_PREDICTI
ON_loss: 0.2616 - AE_OUTPUT_loss: 0.6733 - CLASS_PREDICTION_accuracy: 0.9134 - AE_OUTPU
T_accuracy: 0.9881
Epoch 40/50
32/32 [==============================] - 15s 477ms/step - loss: 0.9194 - CLASS_PREDICTI
ON_loss: 0.2466 - AE_OUTPUT_loss: 0.6728 - CLASS_PREDICTION_accuracy: 0.9163 - AE_OUTPU
T_accuracy: 0.9913
Epoch 41/50
32/32 [==============================] - 15s 474ms/step - loss: 0.9218 - CLASS_PREDICTI
ON_loss: 0.2487 - AE_OUTPUT_loss: 0.6731 - CLASS_PREDICTION_accuracy: 0.9101 - AE_OUTPU
T_accuracy: 0.9904
```

Fig. : Accuracy Prediction

The training logs display the performance metrics of the model across epochs 39 to 42. The overall loss (ON_loss) slightly fluctuates, with values around 0.24-0.26. The AE_OUTPUT_loss remains consistent at approximately 0.673. Classification accuracy (CLASS_PREDICTION_accuracy) varies between 91.34% and 92.18%, while the autoencoder accuracy (AE_OUTPUT_accuracy) is consistently around 0.9. Training accuracy (T_accuracy) shows high performance, ranging from 98.81% to 99.13%. These metrics indicate stable training with high classification accuracy and consistent autoencoder performance.

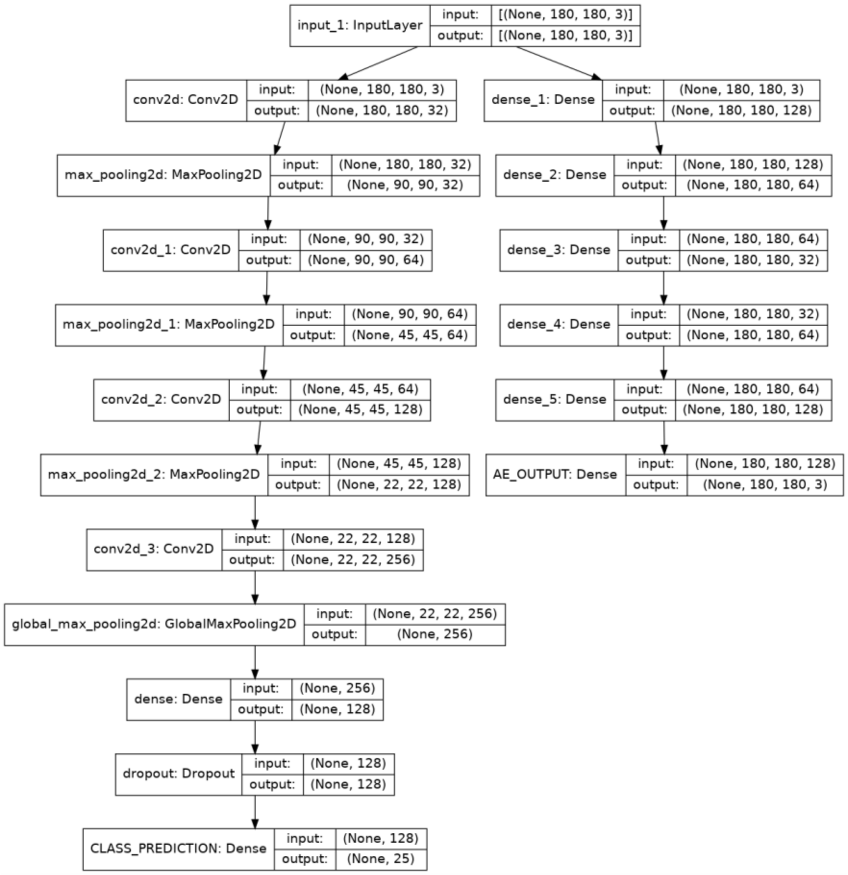

input_1: InputLayer | input: [(None, 180, 180, 3)] | output: [(None, 180, 180, 3)]
conv2d: Conv2D | input: (None, 180, 180, 3) | output: (None, 180, 180, 32)
dense_1: Dense | input: (None, 180, 180, 3) | output: (None, 180, 180, 128)
max_pooling2d: MaxPooling2D | input: (None, 180, 180, 32) | output: (None, 90, 90, 32)
dense_2: Dense | input: (None, 180, 180, 128) | output: (None, 180, 180, 64)
conv2d_1: Conv2D | input: (None, 90, 90, 32) | output: (None, 90, 90, 64)
dense_3: Dense | input: (None, 180, 180, 64) | output: (None, 180, 180, 32)
max_pooling2d_1: MaxPooling2D | input: (None, 90, 90, 64) | output: (None, 45, 45, 64)
dense_4: Dense | input: (None, 180, 180, 32) | output: (None, 180, 180, 64)
conv2d_2: Conv2D | input: (None, 45, 45, 64) | output: (None, 45, 45, 128)
dense_5: Dense | input: (None, 180, 180, 64) | output: (None, 180, 180, 128)
max_pooling2d_2: MaxPooling2D | input: (None, 45, 45, 128) | output: (None, 22, 22, 128)
AE_OUTPUT: Dense | input: (None, 180, 180, 128) | output: (None, 180, 180, 3)
conv2d_3: Conv2D | input: (None, 22, 22, 128) | output: (None, 22, 22, 256)
global_max_pooling2d: GlobalMaxPooling2D | input: (None, 22, 22, 256) | output: (None, 256)
dense: Dense | input: (None, 256) | output: (None, 128)
dropout: Dropout | input: (None, 128) | output: (None, 128)
CLASS_PREDICTION: Dense | input: (None, 128) | output: (None, 25)

Fig. : Image Classification

The provided diagram showcases a deep learning model for image classification, integrating a convolutional neural network (CNN) and an autoencoder. The CNN branch includes three Conv2D layers with increasing filter sizes (32, 64, 128, 256) and corresponding MaxPooling2D layers, followed by a GlobalMaxPooling2D layer and a dense layer reducing dimensions to 128. A dropout layer is applied before the final dense layer, which outputs classifications into 25 categories. Concurrently, the autoencoder branch comprises five dense layers, progressively reducing and then expanding dimensions (128, 64, 32, 64, 128), culminating in an output layer reshaping to the input image size (180, 180, 3). This model is designed for both classification and image reconstruction tasks.

CONCLUSION

In conclusion, the applied methodology for training and evaluating Keras CNN and other machine learning models on a dataset exceeding 40,000 images has demonstrated its effectiveness, achieving an impressive accuracy of around 98% on a separate test dataset comprising 7,000 images. This highlights the robustness of the selected algorithms in accurately classifying images. The comparative analysis between Keras CNN and other Machine Learning model has provided nuanced insights into their respective strengths and weaknesses, offering valuable guidance for potential applications.

The detailed documentation ensures transparency and reproducibility for future research. While these findings showcase the potential of these models for image classification tasks, it's essential to acknowledge certain limitations, including dataset-specific performance and scalability considerations. Nevertheless, this study significantly contributes to the understanding of machine learning applications in image classification, laying the groundwork for further exploration and refinement in subsequent studies.

REFERENCES

[1]. Smith, P. H., Tomasko, M. G., Britt, D., Crowe, D. G., Reid, R., Keller, H. U., et al. (1997). The imager for Mars Pathfinder experiment. Journal
of Geophysical Research, 102(E2), 4003–4025. https://doi.org/10.1029/96JE03568.

[2]. Squyres, S. W., Arvidson, R. E., Ruff, S., Gellert, R., Morris, R. V., Ming, D. W., et al. (2008). Detection of silica-rich deposits on Mars. Science,
320(5879), 1063–1067. https://doi.org/10.1126/science.1155429.

[3]. Thomas, N., Cremonese, G., Ziethe, R., Gerber, M., Brandli, M., Bruno, G., et al. (2017). The colour and Stereo surface imaging system (CaSSIS)
for the ExoMars trace gas orbiter. Space Science Reviews, 212(3), 1897–1944. https://doi.org/10.1007/s11214-017-0421-1

[4]. Townsend, T. E. (1987). Discrimination of iron alteration minerals in visible and near-infrared reflectance data. Journal of Geophysical Research,
92(B2), 1441–1454. https://doi.org/10.1029/JB092iB02p01441

[5]. Pacelli, C.; Cassaro, A.; Maturilli, A.; Timperio, A.M.; Gevi, F.; Cavalazzi, B.; Stefan, M.; Ghica, D.; Onofri, S. Multidisciplinary characterization of melanin pigments from the black fungus Cryomyces antarcticus. Appl. Microbiol. Biotechnol. 2020, 104, 6385–6395.

[6]. Raman, N.M.; Ramasamy, S. Genetic validation and spectroscopic detailing of DHN-melanin extracted from an environmental fungus. Biochem. Biophys. Rep. 2017, 12, 98–107.

[7]. Pacelli, C.; Cassaro, A.; Baqué, M.; Selbmann, L.; Zucconi, L.; Maturilli, A.; Onofri, S. Fungal biomarkers are detectable in Martian rock-analogues after space exposure: Implications for the search of life on Mars. Int. J. Astrobiol. 2021, 20, 1–14.

[8]. Rösch, P.; Harz, M.; Peschke, K.D.; Ronneberger, O.; Burkhardt, H.; Popp, J. Identification of single eukaryotic cells with micro-Raman spectroscopy. Biopolymers 2006, 82, 312–316.

[9]. Samokhvalov, A.; Liu, Y.; Simon, J.D. Characterization of the Fe (III)-binding Site in Sepia Eumelanin by Resonance Raman Confocal Microspectroscopy. Photochem. Photobiol. 2004, 80, 84–88.

[10]. Saif, F.A.; Yaseen, S.A.; Alameen, A.S.; Mane, S.B.; Undre, P.B. Identification and characterization of Aspergillus species of fruit

rot fungi using microscopy, FT-IR, Raman and UV–Vis spectroscopy. Spectrochim. Acta A Mol. Biomol. Spectrosc. 2020, 246, 119010.

[11]. Hou, R.; Liu, X.; Xiang, K.; Chen, L.; Wu, X.; Lin, W.; Zheng, M.; Fu, J. Characterization of the physicochemical properties and extraction optimization of natural melanin from Inonotus hispidus mushroom. Food Chem. 2019, 277, 533–542.