**PROJECT TITLE: Classifying mineral grains using data science. Can data science be used to identify mineral grains, rapidly, accurately, and precisely?**

(CLIENT: Portable Spectral Services; TEAM: G8\_MLP)

**Aim and Background:**

Identifying and counting individual mineral grains composing sand is a vital component of many studies in environment, engineering, mineral exploration, ore processing and the foundation of geo-metallurgy. Typically, this task is performed manually with the drawback of monopolizing both time and resources. Moreover, it requires highly trained personnel with a wealth of knowledge and equipment, such as scanning electron microscopes and optical microscopes. Advances in machine learning and deep learning make it possible to envision the automation of many complex tasks in various fields of science at an accuracy equal to human performance, thereby, avoiding placing human resources into tedious and repetitive tasks, improving time efficiency, and lowering costs. Currently, PSS is using Micro-XRF, which is a scanning technology that rapidly maps the distribution of elements in a mineral grain at the micro scale. Data generated from these scans is routinely used to identify mineral grains present in the sample through a dedicated software program called AMICS. A Spectral Scientist can process up to 40 samples a day through AMICS. Here, we develop deep-learning algorithms to automate the recognition of minerals directly from the grains captured from Micro-XRF, supplying a smart data driven solution for the increased productivity which PSS need.

This proposal introduces an efficient application of Multi-Layer Perceptron’s (MLPs), a class of artificial neural networks, to the task of mineral identification using tabular data derived from mineral samples. The work commences by curating a comprehensive dataset of mineral samples, each accompanied by a set of relevant attributes such as physical properties, chemical composition, and spectral characteristics.

The project's primary aim is to show that MLP-based models can effectively automate the mineral identification process using tabular data, displaying the model's accuracy, robustness, and potential practical applications.

* **Aspects:** Offering a valuable tool for professionals in the fields of geology, material science, and mineral exploration. The research advances the domain by presenting a robust solution for sped up mineral classification and characterization.
* **Versatility:** Model that can autonomously distinguish and classify new minerals based on their unique attributes, even when they have not been encountered before. Unlike traditional methods that rely on predefined mineral classes, this model operates in an open-ended manner, allowing it to explore uncharted mineralogical territory.
* **Time Efficient**: The model automates the process, allowing to quickly identify minerals from the data, saving significant amounts of time, advancing the current limitation of processing 40 samples per day.
* **Reducing Misclassification:** Model can learn complex patterns in data that might not be immediately apparent in the existing implementation. This can lead to improved accuracy in mineral identification, reducing the risk of misclassifications.

**Key Benefits:**

Through the project we aim to supply the following advancements to the existing environment:

* **Pattern Recognition in Data:** The MLP's primary goal is to discern intricate patterns and characteristics embedded in the tabular attributes of various mineral grain by analysing attributes, the model aims to differentiate between mineral types effectively.
* **Scalability and Efficiency:** With the potential to process large tabular datasets efficiently, the MLP can scale to accommodate a diverse range of mineral types. This scalability is crucial for applications involving extensive datasets encountered in geological and material science studies.
* **Generalization to Unseen Data:** Generalize its insights from the training data to accurately classify mineral grains it has not encountered before.
* **Interpretable Insights:** While MLPs are known for their complexity, efforts can be made to interpret the relationships learned by the model, enabling professionals to gain insights into the critical features influencing classification decisions.

**Methods:**

* **Data Preparation:** The tabular data is pre-processed to handle missing values, normalize features, and reduce noise, ensuring a consistent and reliable input for the MLP model. The dataset is partitioned into training, validation, and testing subsets to facilitate model development and assessment. An MLP architecture is designed, comprising input, hidden, and output layers.
* **Model Initialization:** Through a supervised learning paradigm, the model is trained on the training dataset, learning the intricate relationships between input attributes and mineral types.
* **Model Optimization:** Hyperparameter optimization is conducted to fine-tune the model's performance, involving adjustments to learning rates, the number of neurons per layer, and regularization techniques. The validation subset guides these adjustments, preventing overfitting and enhancing generalization capabilities.
* **Model Validation and Testing:** Assess the model's performance using validation data to ensure it meets a predefined level of accuracy. Thoroughly test the model on a separate testing dataset to evaluate its ability to accurately classify minerals that it has not seen during training.
* **Model Performance Evaluation:** Performance metrics such as accuracy, precision, recall, and F1-score are employed to quantitatively evaluate the model's proficiency in accurately classifying minerals. These metrics provide insights into the model's precision and recall rates for different mineral classes.
* **Model Outcome:** The outcomes underscore the viability of MLPs for mineral identification tasks, highlighting competitive performance in discerning diverse mineral types. The model's adaptability across various mineralogical contexts highlights its potential applicability in geological surveys, mineralogical research, and industrial processes.
* **Model Comparision:** Systematically comparing MLP model with existing and already implemented machine learning techniques, to provide a comprehensive assessment of how our approach stands out and contributes to the field of mineral identification.

Thus, given the volume of samples and the data generated per sample, can a data science solution be found that can rapidly, accurately, and precisely identification mineral grains. The consequences of misclassifying a mineral grain can be immeasurable. As a quality control measure, assigning a probability to the mineral grain classification is essential.

**Literature Discussion:**

Most of the literature discussion provides evidence that **Random Forest** has given them maximum accuracy when compared to the available ML techniques, for characterising the mineral grain correctly. In one of the papers, it was discussed that a D**eep Forest** provided them best results over traditional ANN’s and ML techniques. The papers also discuss that the Deep learning techniques proved to be more accurate than ML methods.

I have found a lot of provable citations on Random Forest (few on XGBoost) and ensemble techniques over the internet. Deep learning techniques were heavily used and provided best results in image characterisation. I propose that apart from deep learning technique s we should considerately invest in exploring the cited techniques and then compare the models.

**References:**

* <https://www.sciencedirect.com/science/article/pii/S0098300419301037>
* <https://mdpi.com/2075-163X/12/4/455>
* <https://academic.oup.com/nsr/article/6/1/74/5123737?login=true>
* <https://www.sciencedirect.com/science/article/pii/S0098300421001904>
* <https://chat.openai.com/>
* Aurelian Geron, Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, O’Reilly, 3rd edition, 2022