Classifying Mineral Grains Using Data Science

Can Data Science Be Used to Identify Mineral Grains, Rapidly, Accurately and Precisely?

University of Western Australia

Capstone Project Proposal - Group 08

Bhargava Sai Gopal Kolli (23697081), Christopher Vernon (22732248), Mingxuan Zhou (23745017),

Neha Sharma (23145186), Peter Millitz (23088298), Robbie Baiuo (23028611)

Table of Content:

Aim of the project

The principal aim of the project is to evaluate whether Data Science can be successfully used to identify mineral grains with precision, accuracy and speed.

Background and value proposition

The client, Portable Spectral Services (PSS), established in Perth, Western Australia in 2011, provides in-house sampling services using various spectral technologies including micro-XRF (X-ray Fluorescence). Micro-XRF is a scanning technology used by PSS for elemental analysis in samples at the micrometre scale. The samples in this context are mineral grains of particular interest to the mineral exploration, development and production industry.

Micro-XRF is a non-destructive technique which involves directing a focused X-ray beam onto a sample mineral grain which then interacts with atoms in the grain resulting in the emission of characteristic secondary (or fluorescent) X-rays. The fluorescent radiation is measured in a Silicon Drift Detector (SSD) and analysed by sorting the specific energies of fluorescent photons. The intensity (count rate) of each characteristic radiation is directly related to the elemental abundance present in the sample.

The PSS workflow involves grid scanning an entire sample block approximately the size of a postage stamp with a single layer of mineral grains sieved to roughly the size of a grid cell. The output is a high-resolution two-dimensional qualitative compositional map based on an entire X-ray spectrum from each pixel in the grid. These element maps display the spatial variation and abundance of major, minor and trace elements.[1]

The next step in the workflow involves manual element identification conducted by a spectral specialist to confirm the presence of individual elements and to ensure that the absorption energy peaks are appropriately allocated to each element. The workflow additionally involves processing each X-ray spectrum via proprietary software (AMICS) to identify the mineral phase associated with the sample by comparing it with a comprehensive mineral database. It is this aspect of the workflow which is the target of this proposal.

Current challenges and objectives

A spectral scientist at PSS can currently process up to 40 samples a day using the AMICS processing software to identify mineral phases; PSS would like to increase throughput to around 200 samples per day. This project aims to meet or exceed that objective using automated techniques including Machine Learning (ML) to speed up the identification of mineral grains whilst minimising the level of manual intervention required. In addition, a major drawback with the AMICS software is the inability to add new minerals to its database and it does not attach a probability to its mineral identification. Hence, the proposed solutions will address both issues.

Key benefits to PSS:

- **Reduced data preparation and processing bottlenecks**: Using raw spectral data with minimal pre-processing can help to reduce the time and effort spent on data preparation and processing.
- **Quick and easy retraining of ML models**: Deployed ML models can be quickly and easily retrained to incorporate previously unseen minerals. This overcomes the current drawback with the AMICS software, which requires manual intervention to update the mineral library.
- **Increased throughput:** The proposed techniques can lead to a substantial increase in the throughput of samples analysed, while still maintaining the same level of accuracy and precision.

Expected outcomes of the project

A robust data-driven solution is proposed to identify mineral grains using µ-XRF technology. This solution will use machine learning models to classify mineral grains based on their elemental composition. The models will be trained on a large dataset of µ-XRF spectra, which will allow them to identify a wide variety of minerals. The solution will also assign a validation probability to each classification. This probability will be based on the confidence of the machine learning model in its prediction. A higher validation probability indicates that the model is more confident in its prediction, and therefore less likely to be incorrect.

Data analytics and visualisation outputs

The integrated data analytics outputs or the project will provide standalone interfaces for visualising, interrogating, and classifying mineral grain datasets. These interfaces will enable the following:

- **Visual summaries:** sample statistics including elemental abundance, proportion of useful channels and/or channel ranges, predicted mineral.
- **Model performance graphs**: Graphs will be displayed to show the performance metrics of all machine learning models including accuracy, precision, recall, F1-score and prediction probability.

Presentation, reporting and code

All work will be recorded and stored in a shared GitHub repository. This includes all original and processed data, data exploration outputs, statistical data analysis and models. All code will be thoroughly documented. Meetings with the client will be scheduled on a regular basis to update the client on the team's progress, discuss key aspects of the project, solicit feedback, and maintain the team's focus with respect to the client's expectations. The attached Gantt chart (Appendix 1) provides a detailed schedule of proposed meetings with the client.

Methods

Task 1: Exploratory data analysis and feature selection

The dataset is anticipated to consist of several samples, in tabular format, each sample comprising up to 8000 rows (representing individual spectra) and 4096 columns (channels) which record the radiation count rates in specific energy windows. A significant number of columns will contain zero counts or will be missing altogether. This makes for a very sparse and high-dimensional dataset. Moreover, the dataset is expected to be imbalanced. To illustrate, an example dataset supplied by PSS showed that a single mineral class constituted 75% of the data while 17 other categories made up the remaining 25%.

The proposed solutions will focus on those techniques that can meet the challenges of high-dimensional, sparse and imbalanced data input. Feature importance assessment, Sparse Principal Components Analysis (SPCA) and channel aggregation can be employed to address the issue of high-dimensionality and data sparsity. Methods to mitigate the effect of data imbalance will be investigated, likely requiring a combination of methods including stratified sampling, variable weight reallocation and data partitioning.

This phase will also involve low level processing of the dataset in preparation for the modelling phase e.g. imputing missing values, removing unused channels, data standardisation and randomly splitting the dataset into training and test sets.

Task 2: Experimentation and Data Science solutions

Our team conducted an initial study to compare various models that were capable of handling high-dimensional sparse tabular datasets. The study concluded that both tree-based machine learning models and specific deep learning models tailored for tabular data processing exhibited high levels of accuracy. As a result, the team have decided to employ a range of techniques including those mentioned in the study, namely:

- Tree-based machine learning methods which use hierarchical structures to make predictions, e.g., Random Forest[2] and XGBoost[3]
- Deep Learning algorithms including MLP and TabNet[4,7]
- Supervised of Logistic Regression with L1-Regularisation.[5,6]

It is anticipated that after all models have been evaluated, they will be considered for use in combination to determine if classification performance can be further boosted using Stacking and/or Hard/Soft Voting ensemble methods.

Task 3: Modelling, evaluation, optimisation and results presentation

This phase will involve training various models using the nominated algorithms. Modelling will typically involve an iterative process of hyperparameter selection, fine-tuning, model training and evaluation. Cross-validation techniques will be utilised where possible to determine best plausible parameters/settings and in model prediction performance. Each model will also be evaluated on its generalisation performance on the previously unseen test set.

Selection criteria will place emphasis on the client's requirement that the classifier should return both high precision (actual positives classifications) and high recall (positive classifications actually correct). This can be summarised by a single metric, the F-1 score. Models will also be evaluated on a range of other performance metrics. Results will be communicated to stakeholders via a report and presentation which will be summarised in the form of various tabular or graphical formats, for example, a confusion matrix (aka error matrix), ROC curve (binary classifier diagnostic) and precision vs recall plot, to name a few.

Project deliverable timeline & costs

The project was divided into four major tasks: project initiation, proposal, execution, and deliverables.

- **Project initiation** involved getting to know each team member's strengths and capabilities.
- **Proposal stage,** each team member was asked to provide a review of at least two relevant academic papers in literature relevant to the project.
- **The execution stage** involved implementing the best models based on the literature review and preliminary investigations. This stage also included fine-tuning models and investigating deployment.
- **The delivery stage** involved implementing the model to the client environment, writing the final report, and preparing documentation and a quick reference guide for client use.

At the end of each stage, there is a client meeting to showcase the team's progress and answer any questions. The project deliverables timeline and estimated data scientist hours can be found in Appendix 1.

References

[1] Portable Spectral Services. Micro Chemistry & Mineralogy. Retrieved August 24, 2023, from

[2] Li, C., Wang, D., & Kong, L. (2021). Application of Machine Learning Techniques in Mineral Classification for Scanning Electron Microscopy - Energy Dispersive X-Ray Spectroscopy (SEM-EDS) Images. Journal of Petroleum Science and Engineering, 200, 108178. https://doi.org/10.1016/j.petrol.2020.108178

[3] Ziegler, G. (2019). Multiclass & Multilabel Classification with XGBoost. Medium. Retrieved from https://gabrielziegler3.medium.com/multiclass-multilabel-classification-with-xgboost-66195e4d9f2d

[4] Borisov, V., Leemann, T., Sessler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/tnnls.2022.3229161

[5] Chen, Y., & Zhao, Q. (2021). Mineral exploration targeting by combination of recursive indicator elimination with the ℓ2-regularization logistic regression based on geochemical data. Ore Geology Reviews, 135, 104213. https://doi.org/https://doi.org/10.101/j.oregeorev.2021.104213

[6] Yu, D., Lee, S. J., Lee, W. J., Kim, S. C., Lim, J., & Kwon, S. W. (2015). Classification of spectral data using fused lasso logistic regression. Chemometrics and Intelligent Laboratory Systems, 142,70-77. https://doi.org/https://doi.org/10.1016/j.chemolab.2015.01.006

[7] W. M. Brown,T. D. Gedeon,D. I. Groves &R. G. Barnes (2010). Artificial neural networks: A new method for mineral prospectivity mapping, 757-770. https://doi.org/https://doi.org/10.1016/j.chemolab.2015.01.006

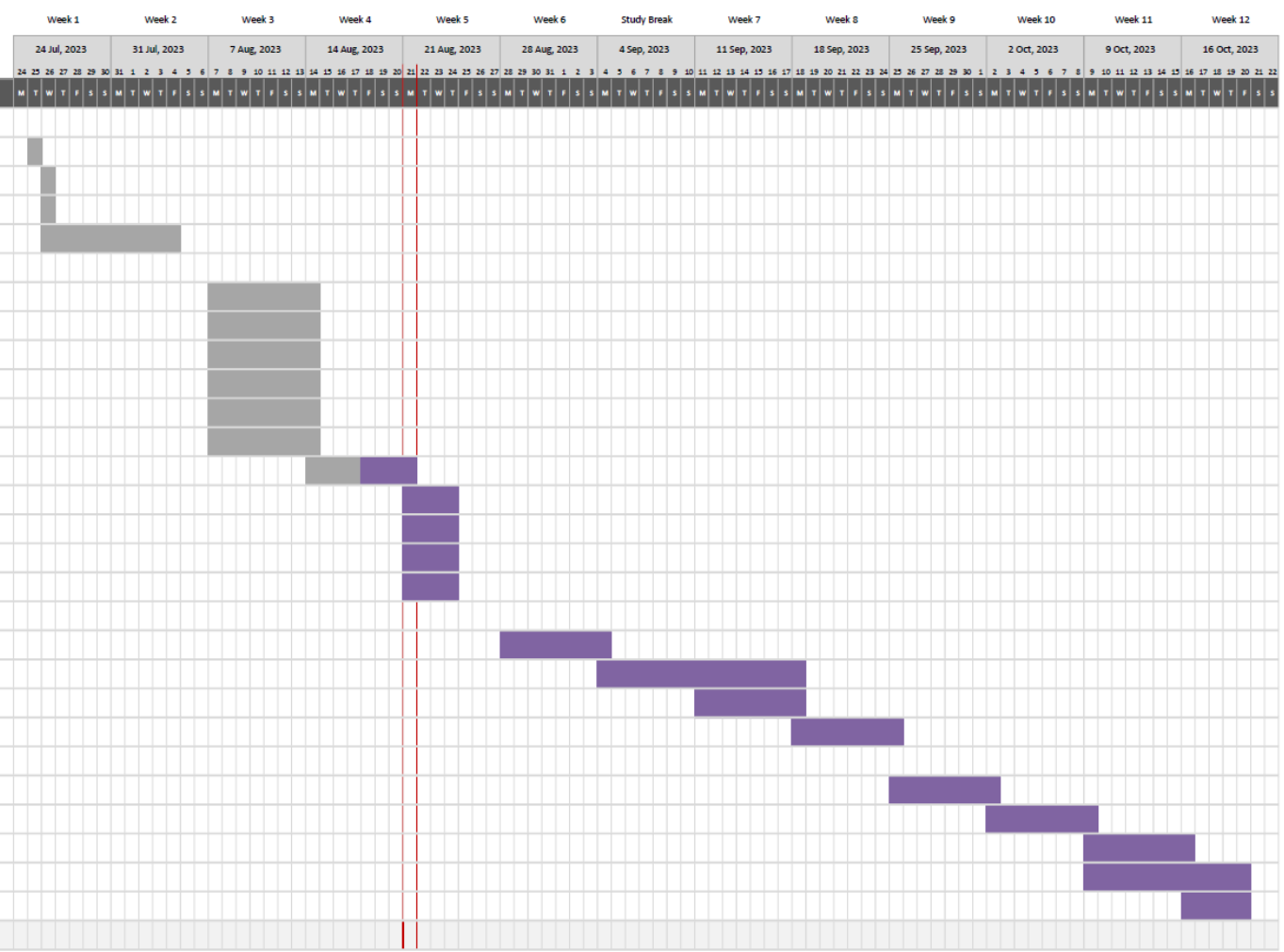
Appendix 1: Project deliverables timeline and estimated hours

CITS5553 Data Capstone Project: Mineral Grains Identification Using Data Science

Group 8
Compiled by: Robbie

Project Start: Mon, 24/7/2023
Display Week: 1

TASK	ASSIGNED TO	PROGRESS	HOURS	START	END
Project Initiation					
Setup MS Teams Comm Channel	Peter/Prof.	100%	1	25/07/23	25/07/23
First Team Meeting	Group	100%	1	26/07/23	26/07/23
Setup GitHub Repo	Ming	100%	1	26/07/23	26/07/23
Client Initiation Meeting (1st Meeting)	Peter	100%	1	26/07/23	4/08/23
Project Proposal					
Literature & Feasibility Review: Tab Net	Ming	100%	6	7/08/23	14/08/23
Literature & Feasibility Review: MLP	Neha	100%	6	7/08/23	14/08/23
Literature & Feasibility Review: Logistic Regression	Peter	100%	6	7/08/23	14/08/23
Literature & Feasibility Review: Random Forest	Chris	100%	6	7/08/23	14/08/23
Literature & Feasibility Review: XGBoost	Robbie	100%	6	7/08/23	14/08/23
Literature & Feasibility Review: LDA	Kolli	100%	6	7/08/23	14/08/23
Prepare Draft Proposal Inputs	Group	50%	24	14/08/23	21/08/23
Review Draft & Update Draft Proposal	Robbie	0%	4	21/08/23	24/08/23
Communicate Proposal to Client	Peter	0%	1	21/08/23	24/08/23
Proposal Review with Client (2nd Meeting)	Group	0%	1	21/08/23	24/08/23
Submit Proposal	Neha	0%	1	21/08/23	24/08/23
Execution					
Develop Individual Models as per Literature Review and Feasibility	Group	0%	36	28/08/23	4/09/23
Invitigate ensemble learning and fine tune final model	Group	0%	48	4/09/23	18/09/23
Increment Show Case to Client (3rd Meeting)	Group	0%	1	11/09/23	18/09/23
Investigate Model Deployment	Group	0%	6	18/09/23	25/09/23
Delivery & Presentation					
Draft Final Report	Group	0%	12	25/09/23	2/10/23
Integrate Best Model into Client Applications	Group	0%	2	2/10/23	9/10/23
Increment Showcase to Client (4th Meeting)	Group	0%	1	9/10/23	16/10/23
Review Final Report and Presentation	Group	0%	6	9/10/23	20/10/23
Submit Final Report	Ming	0%	1	16/10/23	20/10/23
	Total Hours Estimated		197		



Appendix 2: Individual plans and reflections

Robbie Baiuo (23028611) ~ Team Lead

The capstone project is a valuable and challenging task that enhances our theoretical and practical experience on real-world problems. However, working on real-world problems can be difficult and not straight forward. Like many students, as a group we started the project with excitement. But we quickly realised that real-world problems are different from university assignments. Real-world problems are often complex and have no easy solutions. After the first couple of weeks, we realised that we needed to be prepared for complexity. This meant that we needed to think creatively and come up with solutions that were not always obvious or had been studied before. We also realised that we needed to be realistic about our resources. We realised that our client has limited resources and may not be available to help. We also realised that we had a tight deadline and that we needed to manage our time wisely and prioritise our tasks.

Since our project is a multi-label classification problem, we decided to pursue three options to tackle this challenge: machine learning modelling techniques, deep learning, and statistical modelling. At start, we decided to give ourselves a time to study and reflect on what technique would be more suitable by giving each member subject research. This approach helped us to properly conceptualise the solution and the expected challenges we would face. My task was to discover the feasibility of using XGBoost. The papers I read provided a good overview of the use of XGBoost for multi-label classification. I discovered that XGBoost is a powerful algorithm that has been shown to be effective for a variety of multi-label classification problems. It has been widely used by data scientists to achieve state-of-the-art results on many machine learning challenges, but there is a lack of systematic and comprehensive studies on its performance and behaviour. In addition to my individual task as a group member, I took on the challenge of preparing a timeline and deliverable project plan. Having a clear timeline and deliverables made us aware of available resources, tasks, and properly track our progress against plan.

During the first weeks of the project, I have also become aware that leading the capstone project can be challenging considering the different level of expertise among group members, and I must be very careful when assigning tasks. I have also worked out breaking the problem down into smaller chunks can make it seem less daunting and help to make progress on the project. Getting continuous feedback from group members on ideas and progress will help to identify potential problems and make necessary adjustments. Moreover, I should not be afraid to ask for help from team members or mentor if I am struggling with a particular part of the project. By following these points, I am confident that we can increase our chances of success in this project, overcome any obstacles, and gain valuable skills that will be helpful in our future careers.

Peter Millitz (23088298)

Aim & Background

The primary objective is to construct a supervised Logistic Regression classifier Machine Learning (ML) model for mineral identification from micro-XRF spectral data.

The client, Portable Spectral Services (PSS), provides in-house sampling services using micro-XRF (X-ray Fluorescence) scanning technology for elemental analysis in samples at the micrometre scale. The samples are of mineral grains of particular interest in mineral exploration and production. PSS offers a mineral identification service of these samples which is the focus of this project.

Current workflow involves processing X-ray spectra using proprietary software (AMICS) to identify minerals by comparing input spectra with an internal database. Disadvantages of the software include lack of a prediction probability and inability to readily add new minerals to its database. In contrast, a Logistic Regression Classifier inherently provides prediction probabilities and can be rapidly re-trained to recognise new minerals.

The nature of the input data poses three main challenges:

1. The data set is sparse
2. The data set is high-dimensional
3. The data is imbalanced

The data is sparse because each sample consists of up to ~8000 spectra, each with 4096 channels, of which many will record zero counts or no value. The data is high-dimensional because the 4096 channels constitute the predictors in the modelling sense. The data is imbalanced because some minerals are more represented than others. The Logistic Regression ML classifier proposed handles both sparse and high-dimensional input. The imbalance problem will be tackled using a combination of strategies including stratified sampling, data partitioning and appropriate decision rules.

Method

1. Data pre-processing options

- Missing values/zeros replaced with appropriate values
- Delete empty channels and unnecessary columns
- Data standardisation
- Partitioning of the dataset
- Channel aggregation

2. Training & evaluation

The training data set will be randomly split 80:20 into training and test sets respectively, the latter then set aside.

The model will be trained using the Python scikit-learn *LogisticRegression* class[1] with '*L1'* regularisation which can shrink non-influential regression coefficients to zero and handles sparse input.

The modelling process will begin with selection and tuning of hyperparameters. Model predictions will be generated on the training set using 10-fold cross-validation and evaluated by several quantitative and visual metrics.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

A model with suitable performance will then be evaluated on the test set. The results will be reported to the team and depending on outcomes, may be incorporated into an ensemble learner along with other algorithms to boost overall performance.

Personal contribution and reflection:

- Team Leader (weeks 1-3)
- Instigated meetings and team protocols
- Initial design and construction of Gantt chart
- Client/Team mentor liaison
- Literature survey
- Proposal document:
  - Background & Value Proposition
  - Individual proposal
  - final editing

The author believes that a shared leadership model and an equitable decision-making process makes for the most effective project team. Evidently, not everyone agrees. The Team could do with a bit more gender diversity. All said, by the final week of Proposal submission, frustrations were set aside and the team settled to perform effectively.

Neha Sharma (23145186)

When presented with the opportunity to contribute to our university's mining industry project involving mineral grain classification, I was instantly captivated by the potential implications. The fusion of technology and geology presented an exciting challenge, and I was determined to make a substantial contribution. Recognising the project's importance in optimising mining processes, my commitment to the task was resolute from the start. As a team we have decided to explore classification through Machine Learning (ML) and Artificial Neural Network (ANN) techniques. With a dataset of 8000 rows and 4096 features, I turned to the power of Multi-Layer Perceptron (MLP) techniques to address the classification problem. My strategy began with an in-depth exploration of MLP architectures, aiming to harness their capacity to uncover intricate patterns in the dataset. I played a key role in formulating the methods section of our project proposal, outlining a systematic approach to implement sections, majorly focusing on the model implementation part.

The proposed MLP architecture accounted for hidden layer configuration, neurons distribution, and activation functions, aligning them with the specific challenges of mineral grain classification. By contributing to the proposal's methodology, my intent was to lay a solid foundation for our collective efforts. My contribution extends beyond initial planning. Collaborating seamlessly within the team is pivotal. As we advance, my focus remains on data preprocessing – a critical phase in refining the dataset for optimal outcomes. Feature engineering, a facet I'm enthusiastic about, will involve identifying

influential features to enhance classification accuracy. By fostering transparent communication, I aspire to facilitate the exchange of ideas that fuel our progress. The trajectory forward necessitates ongoing experimentation and model refinement. My commitment lies in comprehensive exploration of hyperparameters and optimisation techniques, augmenting the model's efficacy. Through periodic team meetings, our collaborative platform, we will chart our progress, surmount challenges collectively, and steer the project's trajectory.

Moreover, I recognise that the journey towards a successful MLP model implementation requires continuous experimentation and fine-tuning. One of the most significant challenges we face is data preprocessing. The vast dataset demands meticulous feature selection, dimensionality reduction, and normalisation. Ensuring that we retain essential information while eliminating noise is crucial for the model's effectiveness. Additionally, the selection of appropriate MLP architecture parameters is a challenge that demands a nuanced understanding of the problem domain. Balancing the number of hidden layers, neurons per layer, and activation functions requires careful consideration to avoid overfitting or under-fitting.

In conclusion, my involvement in the university project centred on mineral grain classification using MLP techniques has been marked by a profound enthusiasm to merge the realms of geology and cutting-edge technology. Through a continuous exchange of ideas and collective effort, I am confident that we will make significant strides in revolutionising mineral classification within the mining industry.

Christopher Vernon (22732248)

individual Reflection

Portable Spectral Services (PSS) uses micro-XRF technology for elemental analysis of mineral grains, crucial for mineral exploration. The goal is to enhance the mineral identification process using micro-XRF spectral data, aiming to increase sample throughput, reduce manual intervention, and provide prediction probabilities for a more efficient, accurate, and adaptable mineral identification system. This approach seeks to overcome the limitations of the current AMICS software, offering a more streamlined and precise solution for mineral exploration.

In the early stages of the project, the emphasis was on understanding the challenges and objectives of the task at hand. During these foundational meetings, which primarily revolved around discussing the proposal, a proactive role was taken in steering the discussions. This ensured that every meeting was not only productive but also aligned with the project's overarching goals. The collaborative nature of the project was evident with a consistent effort to share insights and knowledge about data science and machine learning methods. I placed importance on assimilating insights from other team members with more experience. By facilitating a two-way flow of information, I ensured that the team's collective expertise was harnessed effectively."

Our team conducted thorough research, looking into three main areas: statistical techniques like Logistic Regression, machine learning methods such as Random Forest, XGBoost, and LightGBM, and deep learning methods including MLP and TabNet. Each team member proposed literature reviews concerning the feasibility of their chosen machine learning method for mineral grain analysis. This

exercise not only fortified our individual understanding of the methods but also provided invaluable insights into how other team members would contribute to the project. The collective focus is to select and further develop the models that demonstrate the best performance. As the project progresses beyond the proposal phase, my primary contribution to the project will be introducing the Random Forest (RF) model to address the challenges posed by the micro-XRF spectral data.

Plan

To begin, I will conduct a in depth preliminary analysis of the micro-XRF spectral data to understand its structure, missing values, and potential outliers. This step is crucial to identify any immediate challenges or requirements for data pre-processing. Given the data's high dimensionality, techniques like Principal Component Analysis (PCA) will be utilised for visualisation and to extract insights into the data's structure. This will also aid in understanding channel correlations and potential feature reduction.

RF provides a natural mechanism for estimating prediction probabilities based on the proportion of trees that vote for a particular class and can rank feature importance. This will enhance model interpretability and provide insights into the most informative channels, in line with the project goals and clients expectations.

Furthermore, techniques like cross-validation will be employed for hyperparameter tuning to optimise the model's performance. The findings will be visualised to effectively communicate the processes and results, encompassing feature importance plots, confusion matrices, and PCA plots. Ultimately, the RF model will be consistently assessed against other machine learning algorithms, facilitating integration, and enabling discussions within the group regarding the performance of alternative models.

Mingxuan Zhou (23745017)

Our primary goal in the project is to classify mineral grains based on spectral data collected from a machine that counts specific elements, with an aim to achieve high accuracy. After consulting with the client, it was understood that the data science solution is expected to assist the client for quality control. This topic is very interesting, offering an opportunity to see how my current knowledge can be applied in an industrial context. While I don't have a background in mineralogy or geology, I see a great potential for further learning and research to aid in the project's success.

The client was welcoming, offering us a tour of their workspace and generously sharing valuable ideas and background information. They were also patient in addressing our team's queries. However, as the project progressed, we faced challenges in communication due to the client's tight schedule. Hopefully, as we move into the implementation phase, we can receive more frequent feedback from the client. This will allow us to pivot as necessary, ensuring we effectively meet the client's requirements and achieve our project goals. Concurrently, our mentor, Chris, provided timely intervention, assisting us in clarifying any uncertainties and provided insightful feedback on our draft proposal.

So far, the project is progressing as planned. The team has completed the proposal and outlined tasks for the upcoming weeks to ensure timely delivery of commitments. A GitHub repository has been set up to facilitate future coding. Weekly meetings are held, and efficient communication is maintained via

Teams for internal communication and Discord for liaising with clients and the mentor. The meetings are organised, with each member receiving specific tasks and clarification regarding individual responsibilities, and addresses about any concerns. My primary contributions include consulting with the client, striving to comprehend the data, conducting preliminary data interrogation, performing a literature review based on current insights, and shortlisting potential models.

In terms of future plans, a significant challenge lies in processing the high-dimensional sparse dataset with excessive zeros and possibly an imbalance in data distribution. Initial attempts with simple models to obtain base accuracy indicate that they struggle to learn from the raw data effectively. To address this, the next steps involve extracting meaningful features, especially emphasising peak detection with the Scipy library, since peaks play an important role in classification. Experimentation with various dimensionality reduction techniques will also be explored. If these strategies do not yield desired results, an alternative approach may involve converting the raw data into images and employing CNNs or image transformers for classification. I am dedicated to adhering to the timelines set in the Gantt Chart, striving towards the project's objectives, ensuring timely updates for the team, and fostering a collaborative environment by seeking and offering assistance as needed.

Bhargava Sai Gopal Kolli (23697081)

Participating in the capstone project has presented an exceptional opportunity, allowing me to become an integral part of a real-time project. This engagement has granted me the privilege to delve into extensive research across diverse domains associated with minerals and geoscience. The practical application of theoretical knowledge in this project is invaluable, as it equips me to navigate authentic scenarios effectively. Furthermore, this experience has been instrumental in enhancing my team management skills and cultivating a professional behaviour in client interactions.

Turning our attention to the collective team performance, there is a palpable enthusiasm among team members to contribute meaningfully to the project's success. We have instituted a rhythm of weekly meetings, where collaborative brainstorming sessions take place to devise innovative approaches for project advancement. These meetings serve as a platform for airing concerns and collaboratively formulating solutions. Each team member is enthusiastic about presenting their unique perspectives, thus delineating a comprehensive scope for the desired project outcome.

My personal engagement within this context has been primarily directed towards data exploration and subsequent visualization. This encompasses the intricate process of data analysis and employs statistical methodologies such as Linear Discriminant Analysis (LDA). The strategic application of LDA for dimensionality reduction is pertinent due to the high-dimensional nature of the dataset. Concurrently, I remain open to contributing my expertise to various other facets of the project, embodying a holistic collaborative spirit.