TabNet Proposal:

Introduction:

The primary objective of this project is to employ data science techniques for the precise classification of mineral grains. The raw data, extracted from spectra, alongside pertinent sample details and grain labels, will form the basis of our analysis. Anticipated to comprise around 4100 columns and 800 rows per sample, the dataset is expected to encompass roughly 1000 samples, as provided by the client.

Based on the current knowledge of the data set, it will be a giant csv, which makes it a massive tabular data. Due to the sufficient data source and the consideration of the data complexity, a deep learning model is proposed to make up for the shortages from traditional models (machine learning models), for DL models are better at handling complex relationships and extracting meaningful features.

TabNet:

TabNet is a deep learning architecture designed specifically for tabular data. It combines the power of neural networks with attention mechanisms to efficiently process structured data. The model employs a novel "sparsemax" activation function, allowing it to perform feature selection while learning to capture relevant interactions between variables. This unique feature makes TabNet particularly effective in handling high-dimensional datasets with both categorical and continuous features.

Benefits:

1.Effective Feature Selection:

TabNet employs an attention-based mechanism that dynamically selects and weighs features during training. This attention mechanism allows the model to focus on informative features while suppressing noise and irrelevant information.

2.Handling High-Dimensional Data:

The TabNet’s attention mechanism effectively performs implicit feature selection by assigning higher attention weights to relevant features. This process can aid in dimensionality reduction, improving the efficiency of the model and potentially enhancing its generalisation.

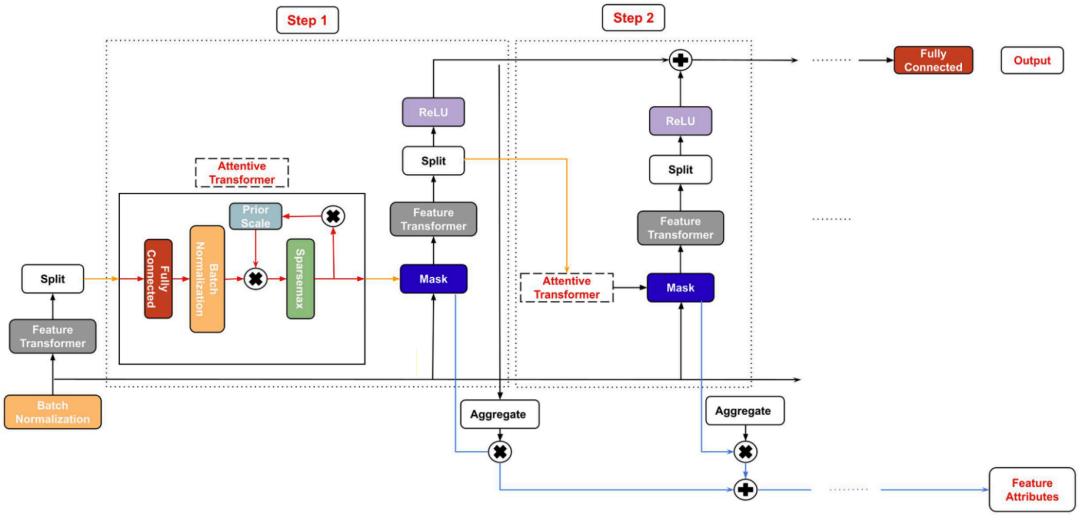
3.Interpretable Decision-Making:

TabNet's attention-based mechanism allows for interpretability by highlighting which features the model focuses on during classification. This transparency enhances interpretability in the model's decisions and enables to validate the model's reasoning behind classifying specific mineral grains.

Implementation Plan:

1. Data Preprocessing and Feature Engineering:
    - Clean raw data, handling missing values and outliers.
    - Normalise or scale the features to ensure consistent ranges.
    - Consider applying dimensionality reduction techniques (PCA) given the high-dimensional nature of the data.
    - TabNet uses a sequential attention mechanism (Sparemax) to choose features for each decision step which does not require much data preprocessing.

2. Model Implementation, Configuration and Tuning:
    Architecture of TabNet:



- Implement the TabNet architecture using a suitable deep learning framework (e.gPyTorch).
- Configure TabNet's hyperparameters, such as decision steps.
- Train the model on the preprocessed data, fine-tuning hyperparameters.

3. Interpretability and Feature Importance Analysis:
    - Local interpretability: TabNet's decision mask, explaining the importance of the input features and how they combined.
    - Global interpretability: Python library Scikit-learn, explaining the amount of contribution of each feature.

4. Model Evaluation:

F1-score, precision, recall, AUC, ROC, Matthews Correlation Coefficient.

Expected Outcomes:

From the literature reviews, the accuracy of Tabnet should be comparable with XGBoost. Depending on the data set, the aim is to reach around 90% of accuracy.

Conclusion:

TabNet emerges as a highly suitable solution for tackling this mineral grain classification challenge. Its adeptness in handling extensive spectral data, coupled with its inherent feature importance analysis and interpretability, perfectly align with the project's goal of accurate classification. By harnessing TabNet's capabilities, the results are not only highly interpretable but also gain valuable insights into data trends with the underlying factors distinguishing mineral grains. Moreover, TabNet is flexible on architecture or can be ensembled with XGBoost for further performance improvement.

Therefore, TabNet is hereby proposed as a potent and promising solution for addressing the challenges of mineral grain classification in this project.