III. Model documentation and write-up

1. My background is in software engineering, with an interest in Machine Learning. I do enjoy participating in machine learning challenges, I have won a number of competitions in the past.

2. It's an interesting and unusual task, a new domain for me. It's a very exciting domain, not every day you can do something which may be used for Mars exploration!

3. I represented the mass spectrogram as a 2D image (temperature vs m/z values) used as an input to CNN, RNN or transformer-based models. The diverse set of preprocessing configurations and models helped to achieve the diverse ensemble of models.

4. Most of the useful relevant images are related to the data pre-processing and are included in the model documentation.

5. Hard to tell, but overall I think the data preprocessing is useful.

6. The dataset is relatively small, models are simple (like resnets34) and the samples are represented as low-resolution 80x100 images, so powerful hardware is not really needed. I have trained a number of different models so it's useful to have access to multiple GPUs.
I used AMD 5950x CPU with a few 2080ti GPUs, 128GB of RAM and Ubuntu 20.04 as an OS.
The submission included the unnecessarily large ensemble of 26 models, trained on 4 folds with 16 samples generated during test time augmentation.
Every model is simple and takes a few hours or less to train, but due to a large number of models, it would take more than a week to train all models sequentially, much faster if trained in parallel on a number of GPUs.

It takes 2 hours to run inference, which can be reduced if necessary by reducing the number of the test time augmentations from 16 to 4, with a very small impact on the score.

7. Not anything I can think about.

8. Only a few notebooks used to play with the different approaches to find the baseline/background.

9. I have used the 4 folds cross validation, with SAM samples shared equally between folds. In addition, I have trained models on the commercial samples data and checked on the SAM samples, but I'd not trust any conclusions with only 16 validation SAM samples (the score on the cross validation was surprisingly good, but it's only 16 samples).
For individual models I have also checked the accuracy score, as the individual models logloss is not as indicative for the performance in the ensemble.
I also checked the correlation between the individual model predictions and between every model and the ensembled mean prediction, to build the most diverse ensemble. I also checked how much the ensemble score changes when new models are added.

I have also tried the GBM based models, the individual models scored slightly better but were much more correlated.

The biggest unknown was the SAM samples, with the very different data distribution compared to commercial samples. Unfortunately with such blind submission, it was hard to validate the models, I expect the performance to be significantly worse on the SAM samples.

10. Not much, I added more models, based on catboost, tabnet and lightgbm. Switching to the focal loss instead of BCE loss seems to help for both individual model and ensemble scores.

11. I'd work to make sure it works well on the SAM data. To collect as much training data as possible. Pre-training on the commercial samples is important, would check if it's practical to augment the commercial samples data to mimic the physics of SAM testbed produced mass spectrograms (like trying to mimic the different ways the background could be distorted, I think it's the better approach comparing to smarted bg subtraction).