Data Quality, Pre-processing and Algorithm Training

HORIZON-CL4-2023-TWIN-TRANSITION-01-02.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Assessing the quality and validity of the generated data & Data pre-processing activities

For the purpose of model training, **COS** has received an experimental **Titanium Surfacing dataset** with a fixed laser configuration. The dataset is transformed into a CSV-style structure so that it can be read consistently by the data pipeline and inspected using standard data tooling.

The first step focuses on assessing whether the dataset is reliable, consistent, and suitable for further analysis. This includes checking the expected data structure, reviewing value ranges, identifying missing information, and confirming the general consistency of the provided experimental results. These checks help ensure that the dataset can be used as a dependable basis for the next modelling activities.

The **pre-processing pipeline** then prepares the dataset for machine learning use by separating the process-related input information from the target quality outcomes. The information is organised according to its type so that suitable preparation steps can be applied before model training.

COS also introduces a small set of physics-informed engineered features, including Power x Speed, Power per Hatch, and Energy Density. These features are generated from the available process inputs and are intended to help the model better capture meaningful relationships in the experimental data - especially where non-linear behaviour may be present.

As part of the pipeline, the data is further prepared through standard transformation and normalisation steps. Certain values that may be more sensitive to outliers are treated accordingly, and the final predicted quality criteria are returned in a structured format that can support further analysis and optimisation.

The work under these tasks turns the raw experimental dataset into a cleaner and more model-ready representation. It also helps avoid artificial performance gains from trivial mappings and keeps the model focused on the most relevant quality criteria for future prediction and optimisation activities.

Development and training of the algorithm

The algorithm development uses the prepared experimental data to predict process quality outcomes and support future optimisation. Building on the earlier data quality and pre-processing work, **COS** is testing how different modelling approaches can learn the relationship between process settings and the resulting quality criteria.

COS is currently working with a **forward modelling approach**. In this setup, the model first learns how the selected process settings affect the expected quality outcomes. An optimisation step can then be used to search for process settings that bring the predicted results closer to the desired targets. This approach is useful because it follows the natural cause-and-effect structure of the process and also allows for adjustments, such as constraints, parameter limits, objective weights, and trade-offs.

For the current **Titanium Surfacing dataset**, **COS** is testing a **Gradient Boosted Trees model** using an 80/20 train-test split. The current configuration uses a maximum tree depth of 4 and a learning rate of 0.05. This provides a practical starting point for evaluating model behaviour while keeping the setup relatively controlled and easy to interpret. The model is tested both as a baseline version and as an enhanced version using the physics-informed engineered features developed during pre-processing.

The comparison shows a clear improvement when the physics-informed engineered features are included. The enhanced version reduces the typical prediction error for **average pocket depth** by around 30.5%, while Root Mean Square Error (RMSE) is reduced by around 53.4%. For **surface roughness**, Mean Absolute Error (MAE) is reduced by around 45.7% and RMSE by around 24.2%. For the **material removal rate**, the model also improves, with MAE reduced by around 22.0% and RMSE by around 25.0%.

These results indicate that the physics-informed engineered features add useful process context to the model. Instead of relying only on the original measured inputs, the model can use additional indicators that better reflect the interaction between key process settings. On the controlled dataset, the engineered model reaches a strong overall performance level, with a global R^2 value of approximately 0.94.

While results are encouraging, more varied data will be needed to ensure the models are more robust across different materials, processes, laser sources, optics, and machining modes.

This marks the next step after data preparation: using the cleaned and enriched dataset to train models, compare modelling strategies, and assess how well the algorithm can predict process results for future optimisation within the **FLASH-IANUS platform**.