



D10.5 Validation results draft model

Version 1.0

Due date	30.06.2026
Dissemination level	PU
Work package title	Single-Operation Machine Learning
Lead editor	Gunnar Mathiason (HiS)
Internal reviewer	ISEND
Status	Final version

*This project has received funding from the European Union under grant
agreement NUMBER —101178721— ProcTwin*

The information and views set out in this document do not necessarily reflect the official opinion of the European Commission. The European Commission does not guarantee the accuracy of the data included in this document. Neither the European Commission nor any person acting on the European Commission's behalf may be held responsible for the use which may be made of the information contained therein.

Authors

Name	Partner
Gunnar Mathiason	HiS
Abed Bakkour	HiS

History

Date	Version	Authors	Description
22.06.2026	0.1	Gunnar Mathiason, Abed Bakkour	Document structure and content overview. Initial texts.
	0.2	José Ignacio Díaz	ISEND review
	1.0	Gunnar Mathiason	Section 4 updated

Abstract

This deliverable describes the validation work for the first draft machine learning models in ProcTwin. The goal of WP10 is to have a first validated data-driven machine learning model, and D10.5 is meant to establish a validation protocol, both for validation in this WP and for coming model validations. In ProcTwin, each subsequent model development has planned validation steps, and for each WP iteration, the validation protocol will be further developed and described as a deliverable.

Two production routes are modeled: wire rolling at GSW, and steel plate rolling and treatment at SSAB. Having these models is a start from where to iteratively develop refined models further on in the ProcTwin project.

Any model will have limitations since it approximates reality under certain assumptions, and the first draft model has many uncertainties involved. Since the models of WP10 are early versions, it is difficult to assess how well they mimic the complex production. D10.5 is a starting point for assessing increasingly larger models, as they are planned for ProcTwin. The validation protocol of this document is intended both to ensure validity of D10.4 and be extensible to cover the next iterations. By establishing a validation operational protocol, coming iteration can be validated in a similar way, regardless of the current level of modelling detail of each model iteration.

Table of contents

Authors	1
History	1
Abstract	2
1. Introduction.....	4
1.1 Scope	4
1.2 Deliverable role with the project.....	4
1.3 Outline	4
2. Validation for GSW wire rolling.....	5
2.1 Regular data and assessment meetings.....	5
2.2 Evaluation model behavior	5
2.3 Next steps	5
3. Validation for SSAB plate rolling.....	7
3.1 Regular data and assessment meetings.....	7
3.2 Evaluation model behavior	7
3.3 Next steps	7
4. Conclusions	9

1. Introduction

1.1 Scope

This document is the fifth deliverable of WP10. It documents the initial generic approach for ensuring the validity of data-driven models, and it is using the first draft machine learning models (D10.4) as an example for how to ensure this. The activities in WP10 aim to produce datasets and models, first for individual machines (operations), emphasizing feature selection, data preprocessing, and integration to enable effective model development, then extended in coming work packaged to include multiple operations. D10.5 describes how to make sure that every model iteration can be validated as good as possible. Coming validation for model deliverables will be based on D10.5, for refinements of validation for increasingly complex models.

1.2 Deliverable role with the project

Many modeling work packages in this project end with a validation step. If a model is not properly validated against for instance external factors, outside of the collected data and modelling assumptions used, there is a risk that this model does not measure up to the actual production situation, such as the model captures an incomplete or incorrect behavior. A number of assumptions are used when modeling complex phenomena such as manufacturing operations. These assumptions cannot easily be optimally decided, or even be transparently known, and need to be assessed by collaborative engagement of all experts involved in the project.

To be able to develop valid complex models such as in this project, there needs to be a procedure or protocol for how to ensure that those assumptions are not implicit or otherwise unknown. The approach taken is based on a common understanding and sharing of knowledge between project partners, such that explicit data and modeling decisions can be taken with insight of all experts. Our main protocol is based on regular deep-knowledge discussions about processes, data collection, and modeling assumptions, examined in collaboration with relevant project partners.

D10.5 results in the following:

- A description of validation steps taken up until deliverable D10.4.
- A proposal for a generic validation protocol.
- Steps taken at the current stage in the project to implement this protocol.

1.3 Outline

The document describes the following:

- Validation of the draft data-driven GSW wire rolling model.
- Validation of the draft data-driven SSAB plate rolling model.
- Conclusions and establishing a generic validation protocol.

2. Validation for GSW wire rolling

To motivate the proposed validation protocols, we describe how validation was used in WP10 for the GSW wire rolling route.

One example of a deliberate collaboration choice is the data used for data-driven machine learning. Several data sets have been considered for data-driven modelling of the rolling route. At this stage, there was a choice between using a raw time-series copy of the most detailed-level production database or using aggregated time-series from an overall server where time-series were already aggregated. An advantage with detailed time-series data would be to be able to connect sensor measurements to physically correlated output wire positions. However, traceability from sensors to wire positions was not available in this data set. The decision for this stage was to use aggregated data, where traceability was already established. It is noteworthy that this collaborative choice could only be found and decided with all GSW case partners participating.

For the GSW part of the deliverable D10.5, we used two complementary approaches for validating data and models developed in WP10:

2.1 Regular data and assessment meetings

In a series of meetings, the details of the chosen data set have been discussed in depth. For instance, specific pre-processing of the data was discussed. Some pre-processing steps were rejected while others were decided to be used, based on how well they would support model learning. Data choices were decided in collaboration. Some directly missing data parameters were decided to be extracted from the other collected parameters. Such decisions cannot be made by the modeler in isolation. Only collaborative discussions about data pre-processing could reveal such issues. Thus, regular and detailed discussions were used to ensure data validation.

2.2 Evaluation model behavior

At this stage, a few basic non-linear supervised learning models were used. The purpose was to evaluate how well the selected data could predict the target (the measured coil defects). The point was not just to find the best algorithm, but to see how the prediction differs over basic models for this particular data. For the GSW case, two different data periods were chosen as alternatives, and the resulting predictions were compared.

The ranking of how influential parameters were for the target was discussed with the process experts in the coil production team. Currently, the interpretation of ranked influential parameters to understand production causality is difficult, so this approach needs to be complemented. At this point of the project, it is however the single way to validate a model's qualitative performance. Future model validations still need to be based on an assessment of model predictions, as a collaboration between modelers and process experts, which jointly can justify or reject the model behavior.

Complementing parameter rankings to clarify interpretation by process experts needs to be elaborated. Currently we see two ways this can be done 1) Examples of typical correlations for each top-ranked variable. Rankings can be assessed if there are some other statistical descriptions of the correlation based on examples of the correlation ranked high by a model, and the reason is that the number of possible predictors in the dataset is very large, such that manual evaluation is hardly impossible even with extensive process knowledge. 2) Even deeper discussions and hypothesis formation and testing of plausibility of ranking lists is needed.

2.3 Next steps

For the next project stages, we will develop these parts of a more solid validation protocol.

- Data refinement, for further extraction of possible features that are candidates as good predictors for coil defects. This is dependent on the process experts' advice.
- Model refinement, for clearer ranking agreement between the different models that are trained on the available data.
- Ablation study, to investigate the influence of adding data from multiple machines.

3. Validation for SSAB plate rolling

The SSAB heavy-plate route covers five sequential operations, from reheating and hot rolling through blasting, quenching and cold leveling, with the post-leveling plate flatness as the prediction target. Each operation was first modeled individually, and a cross-process model was then trained on the joined data from all operations. The dominant challenge at this stage was the reconciliation of many heterogeneous source tables into a single consistent data set. Validation ran along two lines: collaborative work on the data, and evaluation of how the models behave.

3.1 Regular data and assessment meetings

The largest single validation effort was the integration of the source data, carried out in close and repeated discussion with the SSAB process experts. The data is spread across many separate tables, with differing identifier schemes and time bases, and the records were reconciled following the experts' documented join strategy. Part of this reconciliation was eased by preparation on the SSAB side, where the source data had been adapted to join more reliably than in its raw database form. Several points could only be resolved through these discussions; two are described here as examples. A logging convention in the production systems made one upstream gauge appear to cover only part of the production; the experts explained the convention and provided the key needed to attribute the measurement correctly, which raised its coverage to the full plate population. The orientation of the spatial flatness grid, and which physical plate edge each measurement position corresponds to, was likewise worked out through the experts' analysis of concrete example plates. Alongside these joint decisions, the modeling team made its own choices to keep the models sound: a temporal filter so that only causally valid measurements enter the target, explicit handling of measurements from partially covering equipment rather than discarding them, and the removal of identifier and timestamp fields that a model could otherwise use as shortcuts instead of the physics.

3.2 Evaluation model behavior

For each operation that produces a plate-level measurement, several data-driven algorithms were compared, from regularized linear baselines to gradient-boosted trees, together with a per-pass sequence model for the hot-rolling data. The aim of comparing several algorithms was not to crown a single winner but to see how stable the prediction is across model families and across the alternative preparations of the target. Each model yields a ranked list of its most influential features (the built-in importance for the tree models and a permutation-based ranking for the sequence model). A ranked feature list is hard to confirm without process knowledge, so checking whether the most influential variables are physically sensible is a validation step still to be carried out with the experts. The joined data also supports a cumulative analysis: refitting the model as each operation's data is added measures how much each operation contributes, which assesses the value of combining operations rather than the accuracy of any single model.

Beyond this in-distribution evaluation, the models were tested for their ability to generalize across production periods. In addition to a standard cross-validation, in which the folds are drawn from all three months pooled together so that every period appears in both training and test, each production period was held out in turn and predicted by a model trained only on the other periods. This showed that the current draft models generally do not yet transfer to an unseen period: a model can appear accurate in mixed cross-validation and still fail on a held-out period. We regard this as a central validation finding, because it shows that in-distribution accuracy on its own is not a sufficient validation criterion, and it directly informs the generic recommendation below.

3.3 Next steps

For the next project stages, we will develop these components of the validation protocol.

- Data refinement, for further extraction of candidate predictors from the rolling pass records and the cold-leveling cylinder data, in continued discussion with the SSAB process experts.
- Validation of the feature rankings against process knowledge, so that the most influential variables can be confirmed or rejected.
- Integration of partner physics-based virtual sensor data to provide outputs for the operations that the line data cannot, beginning with the post-quench plate shape and mid-thickness microstructure from the quench simulation, once it is available for our plates.
- Investigation of the cross-period generalization gap, together with adoption of an out-of-period holdout, in which a model is evaluated on an unseen production period, as a standing check repeated for each model iteration and as further production periods become available.

4. Conclusions

Based on the experiences from validation of data and models in WP10, we propose the following components as a generic data and model validation protocol, also to be used as basis for future data and model validations:

- Insightful and interactive preparation of both data and model design parameters.
- Frequent development interactions between all experts, typically as regular work meetings, discussing the details of the choices made during development.
- A collaborative assessment of how models perform. This assessment is not only about model accuracy, but also how the model can be explained in relation to the actual production situation. For such assessment, tight collaboration between the models and the process experts is required.
- A test of whether the model generalizes beyond the data it was fitted on. In-distribution cross-validation is necessary but not sufficient: each development iteration should not only be evaluated on a held-out production within the period, but also out-of-period performance. This out-of-period test should be repeated for each model iteration and as further production periods become available.
- The current protocol assumes a data-driven modelling approach that can be utilized as a basis for model validation in general later on in the project.

The iterative protocol allows data and model to be stepwise refined. Each iteration ends with a validation conclusion defining the degree of validity as agreed by the collaborating experts. Whether external unseen data is used in the final validation assessment will depend on the availability of such data at that stage.

Each validation deliverable document (such as D10.5) should elaborate on how validation was done during and at the end of the activity leading to a deliverable (as described in section 2 and 3 above), with transparent validation assessment and statement.

Figure 1. Validation protocol overview and iterations.

