Markku Ohenoja*, Aki Sorsa and Mika Ruusunen

Interpretable machine learning for transparent industrial applications

Abstract: Decision-making needs typically be transparent and this applies also to model-based design and operation in process industries. User trust is also important to ensure that the developed mathematical models are accepted by practitioners. However, modern artificial intelligence approaches can result in overcomplex models with limited interpretability. This research discusses the topic of interpretable models and studies the applicability of one method in several case studies related to process automation and process engineering. The approach enhances model interpretability and ensures that models can be effectively integrated and maintained with minimal disruption.

Keywords: model complexity, multiple linear regression, genetic algorithms, hydrothermal liquefaction

*Corresponding Author: Markku Ohenoja: University of Oulu, Finland, E-mail: markku.ohenoja@oulu.fi

1 Background and aims

Mathematical models in process automation and process engineering must be explainable for several reasons. First of all, understandable models can result in significantly higher user trust. Secondly, more transparent models provide clear insights into predictions and ensure that the decision-making process is understood and accepted. When models are explainable, it is easier to identify and address errors or biases, thus ensuring that the model's decisions are fair and unbiased.

The model structures generated with artificial intelligence (AI) tools, such as machine learning (ML) or deep learning can be very complex. Explainability is key for responsible AI development, ensuring transparency, accountability, and trustworthiness. Interpretable insights provided by explainable models offer valuable information about data relationships and patterns, which can be particularly useful for domain experts. Additionally, explainable models facilitate collaboration between data scientists and domain experts, leading to more effective and impactful solutions.

According to (Hu et al., 2021), although model complexity is a fundamental problem in deep learning, it still has been in the infant stage. Relevant tools and practices for traceability of AI models to elaborate reproducible or repeatable data analysis have been reviewed by (Mora-Cantallops et al., 2021). Among the practical approaches for interpretable models, LIME (local interpretable model-agnostic explanations) algorithm has been used for different applications, such as wind power forecasting (Yang et al., 2023). Generic methods for more interpretable ML involve using linear, surrogate or rule-based models, and sensitivity analysis, among others.

This work highlights the applicability of a studied ML method being able to produce such interpretable models for process engineering applications. The methodology has been applied in several case studies and this presentation collects the results and also uses the method in a novel problem related to production of green aviation biofuels.

2 Material and methods

The ML method used here is based on Genetic Algorithms (GA) and results in a multiple linear regression model (MLR) with respect to features, but non-linear with respect to real measurements. The algorithm selects variables and performs functional transformations from a pre-defined list, creating features. It then selects mathematical operations (add, multiply, divide) between the features. Thus, GA creates many candidate model structures, evaluates their performance, and uses GA operations (reproduction, mutation, elitism) to generate new populations until convergence or calculation budget is reached. This approach helps build an easily interpretable model and adds user trust. The implementation of the algorithm is described in detail in (Ohenoja et al., 2018) and (Sorsa et al., 2013).

In this work, the method is applied to a dataset consisting of quality control system measurements of a paperboard production line, and to a dataset comprising the reaction conditions and biocrude yield

in hydrothermal liquefaction of microalgae. The latter dataset can be found in (Mordechai Koskas et al., 2023) and was in this study screened to Chlorella Vulgaris algae species.

3 Results

In (Ohenoja et al., 2018), the proposed ML approach was successfully applied to model the polarization curve of fuel cells in different operation conditions. In (Sorsa et al., 2013), it was applied to Barkhausen noise data set to predict stress in steel samples. Both studies demonstrated the method's capability for small datasets and effectiveness in capturing complex relationships within the data.

Regarding the case of predicting the reel dry weight in a two-ply board machine, the dataset contained 64 selected measurements and over 500,000 data points. The mean absolute percentage error (MAPE) of 2.89% was achieved with the proposed ML approach. The identified model utilized five explanatory variables, which were related to the wire speed, headbox ash content, and water and pulp consistencies. The selected variables are also intuitively closely linked to the produced paper grade, and thus the reel dry weight, supporting the successful variable selection of interpretable model. The final selected variable (or its feature) was the pressure filter reject flow rate, having a less intuitive effect to the predicted variable.

Table 1. Model performance metrics in HTL case. Error metrics are calculated from normalized values using test data set (n=17).

Model	RMSE [-]	R [-]	MAPE [%]
Kinetic	0.2969	-0.24	81.9
MLR	0.2072	0.63	64.1
PLSR	0.2069	0.63	64.0
PCR	0.2066	0.61	26.7
ANN	0.1930	0.71	65.3
GA	0.1891	0.73	46.0

The second case study aimed to predict the biocrude yield (Y_{BC}) in hydrothermal liquefaction (HTL) process, an intermediate step in production of green aviation biofuels. Several modeling approaches were tested, such as the kinetic model in (Sheehan & Savage, 2017), and typical ML approaches including MLR, PLSR, PCR and ANN. The model performance metrics are given in Table 1. The GA-based approach outperformed other methods tested in terms of correlation coefficient (R) and root mean squared error (RMSE) and had the second lowest MAPE value. The final model structure with four explanatory variables (residence time RT, feed biomass protein content, F_P, temperature T, feed biomass carbohydrate content F_C) was:

$$Y_{BC} = a_0 + a_1 RT^{-1} + a_2 \frac{\sqrt{F_P}}{T} + a_3 T^3 + a_4 F_C^3.$$

4 Conclusions

The proposed GA-based ML method identified usable features and interpretable models in earlier published case studies and in two new cases presented in this paper. Such models can gain user trust due to their simple model structure. In addition, regression models are easy to implement in industrial automation systems and require less complex maintenance. This approach not only enhances model interpretability but also ensures that models can be effectively integrated into existing systems with minimal disruption.

Future work could explore the application of this method to other domains and further refine the algorithm to improve its efficiency and accuracy. Additionally, incorporating user feedback into the model development process could lead to even more robust and trustworthy models.

5 References

Hu et al. (2021). Model complexity of deep learning: A survey. Knowledge and Information Systems, 63(10), 2585–2619.

Mora-Cantallops et al. (2021). Traceability for Trustworthy AI: A Review of Models and Tools. Big Data and Cognitive Computing, 5(2).

Mordechai Koskas et al. (2023). Process simulation for mass balance of continuous biomass hydrothermal liquefaction with reaction kinetics. Energy Conversion and Management: X, 20, 100477.

Ohenoja et al. (2018). Model Structure Optimization for Fuel Cell Polarization Curves. Computers, 7(4).

Sheehan & Savage (2017). Modeling the effects of microalga biochemical content on the kinetics and biocrude yields from hydrothermal liquefaction. Bioresource Technology, 239, 144–150.

Sorsa et al. (2013). An Attempt to Find an Empirical Model between Barkhausen Noise and Stress. Materials Science Forum, 768–769, 209–216.

Yang et al. (2023). Investigating black-box model for wind power forecasting using local interpretable model-agnostic explanations algorithm: Why should a model be trusted? CSEE Journal of Power and Energy Systems, 1–14.