HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230110

XAIProcessLens: A Counterfactual-Based Dashboard for Explainable AI in Process Industries

Gianluca Manca[a,1], Nilavra Bhattacharya[b], Sylvia Maczey[a], Dawid Ziobro[c], Emmanuel Brorsson[c], and Magnus Bång[d]

[a]*Industrial AI, ABB Corporate Research Center, Ladenburg, Germany*
[b]*IX Lab, University of Texas at Austin, USA*
[c]*User Experience, ABB Corporate Research Center, Vasteras, Sweden*
[d]*Department of Computer and Information Science, Linköping University, Sweden*

ORCiD ID: Gianluca Manca https://orcid.org/0000-0001-5951-8590, Nilavra Bhattacharya https://orcid.org/0000-0001-7864-7726, Dawid Ziobro https://orcid.org/0000-0002-7947-772X, Emmanuel Brorsson https://orcid.org/0000-0003-4238-5976, Magnus Bång https://orcid.org/0000-0003-3606-8579

Abstract. We present a novel counterfactual-based dashboard for explainable artificial intelligence (XAI) in process industries, aimed at enhancing the understanding and adoption of machine learning (ML) models by providing transparency, explainability, and performance evaluation. Our dashboard comprises two modules: a statistical analysis module for data visualization and model performance assessment, and an XAI module for exploring counterfactual explanations at varying levels of abstraction. Through a case study of an industrial batch process, we demonstrate the dashboard's applicability and potential to increase trust in ML models among stakeholders, paving the way for confident deployment in process industries.

Keywords. Counterfactual explanations, dashboard, explainable AI, human-in-the-loop, machine learning, process industries

Extended Abstract

With the advent of advanced control systems, the complexity of modern industrial processes, such as chemical and petrochemical plants, has increased significantly, making it challenging for human analysts to comprehend the underlying dynamics and dependencies efficiently. As a result, machine learning (ML) and artificial intelligence (AI) have become crucial tools in process industries to optimize plant operations and increase efficiency [1,2]. However, the lack of transparency, explainability, and interpretability of black-box ML models often poses a challenge to their adoption by human operators who need to understand the underlying data and reasoning behind the model's predictions [3]. To address this challenge, we present a novel dashboard[2] that

[1] Corresponding Author: Gianluca Manca, E-mail address: gianluca.manca@de.abb.com.
[2] https://explain-project.eu/blog1/

provides a human-machine interface for analyzing and explaining ML models in process industries. Our dashboard facilitates the model's performance evaluation, offering the user, i.e., ML engineers and data scientists, the ability to understand the model's behavior and output via interactive visualization and explanation. The innovation in our approach stems from incorporating counterfactual explanations, addressing a gap in existing solutions and offering a tailored experience for process industries.

We demonstrate the applicability of our approach using a case study of an industrial batch process involving a single chemical reactor, where three distinct solid educts, cooling water, and steam are utilized across a 14-phase procedure, including educt addition, centrifugation, reaction, cooling, and material transfer [4]. To classify the different batch phases, we utilized a state-of-the-art convolutional kernel-based time series classification method, as described in [5]. The 14 phases are distinguished based on their characteristic dynamics, which are derived from the time series trends of five different process variables across a 20-minute sliding window. These variables are continuous sensor readings, including the reactor level and its steam inlet flow [4].

Our dashboard is composed of two modules: a statistical analysis module and an explainable AI (XAI) module. The statistical analysis module provides users an interface to explore and visualize historical data, allowing them to better understand the process and its dynamics. Through examining time trends in historical batches and phases and utilizing various tools, such as data selection, filtering, and trend visualization, one can obtain valuable insights into the data used to train and test the ML model.

Within the statistical analysis module, users can further analyze the performance of the ML model using various tools. These include a confusion matrix [6], which shows how well the model discriminates between different classes, as well as different performance metrics such as the balanced accuracy or the Matthew correlation coefficient [7]. From this analysis, one can drill down to specific examples, allowing for a more in-depth understanding of certain classes. These capabilities permit evaluation of the performance of the ML model and identify areas where it may require improvement, for example by adjusting its parameters or retraining it with additional data.

In recent years, explaining ML models and their outputs has been an active research topic [8]. XAI methods have been developed that aim at providing human-interpretable explanations of the predictions made by black-box ML models [3,8]. Among various XAI methods, counterfactual explanations have been selected for our approach due to their ability to present hypothetical inputs that would result in different prediction outcomes [9]. This approach supports understanding of the model's behavior by allowing users to explore and comprehend the relationships between input features and the resulting predictions, thus providing insights into the model's decision-making process. Our dashboard's XAI module allows to explore these counterfactuals at different levels of abstraction, including high-, medium-, and low-level abstraction explanations. This capability enables users to understand how changes in inputs could lead to different prediction outcomes, thereby enhancing the transparency of the ML model.

The high-level abstraction explanations in our XAI module provide an overview of the most crucial features influencing the predictions of the ML model over all classes. Medium-level abstraction explanations, on the other hand, focus on the features that the model uses to distinguish between any two phases. Finally, low-level abstraction explanations offer a detailed look at how the model can be confused and the prediction outcome altered by allowing the user to interactively generate counterfactuals. These different levels of explanations might help to gain a better understanding of how the ML model works and the features it relies on for making predictions.

In this work, we propose a dashboard that provides a valuable off-line tool for ML engineers and data scientists in the process industries to better understand ML models, facilitating system improvements and post-analysis of ML behavior to comprehend the underlying decision-making process, which in turn can enhance real-time applications for operators when the models are deployed in the field. By providing insight into the underlying data, model performance, and explanations for individual predictions as well as overall model behavior, our dashboard can increase trust and facilitate the adoption of these models by allowing ML engineers and data scientists to better comprehend the decision-making process and justify their models to stakeholders, ultimately leading to a more confident deployment of ML solutions in process industries. Future work will involve conducting user tests to evaluate the effectiveness of our dashboard and incorporating other explainer components, as well as extending it to additional ML applications in the process industries.

Acknowledgement

The project on which this article is based was funded by the German Federal Ministry of Education and Research (BMBF) under the grant number 01IS22030. The responsibility for the content of this publication lies with the authors.

References

[1] Klöpper B., Dix M., Siddapura D., Taverne L.T. Integrated search for heterogeneous data in process industry applications. Proceedings of the IEEE 14th International Conference on Industrial Informatics; 2016 Jul. 19–21; Poitiers, France; pp. 1306–1311. doi: 10.1109/INDIN.2016.7819369.

[2] Manca G., Dix M., Fay A. Clustering of similar historical alarm subsequences in industrial control systems using alarm series and characteristic coactivations. IEEE Access. 2021 Nov; vol. 9; pp. 154965–154974. doi: 10.1109/ACCESS.2021.3128695.

[3] Kotriwala A.M., Klöpper B., Dix M., Gopalakrishnan G., Ziobro D., Potschka A. XAI for operations in the process industry. Proceedings of the AAAI Spring Symposium Combining Machine Learning and Knowledge Engineering; 2021 Mar; Palo Alto, CA, USA.

[4] Vicente M.L.C., Granjo J.F.O., Tan R., Bähner F.D. A benchmark model to generate batch process data for machine learning testing and comparison. Computer Aided Chemical Engineering. 2022 Aug; vol. 51; pp. 217–222. doi: 10.1016/B978-0-323-95879-0.50037-0.

[5] Tan C.W., Dempster A., Bergmeir C., Webb G.I. MultiRocket. Data Mining and Knowledge Discovery. 2022 Jun; vol. 36; pp. 1623–1646. doi: 10.1007/s10618-022-00844-1.

[6] Stehman S.V. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment. 1997; vol. 62; no. 1; pp. 77–89. doi:10.1016/S0034-4257(97)00083-7.

[7] Matthews B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta–Protein Structure. 1975; vol. 405; no. 2; pp. 442–451. doi:10.1016/0005-2795(75)90109-9.

[8] Theissler A., Spinnato F., Schlegel U., Guidotti R. Explainable AI for time series classification. IEEE Access. 2022 Sep; vol. 10; pp. 100700–100724. doi: 10.1109/ACCESS.2022.3207765.

[9] Ates E., Aksar B., Leung V.J., Coskun A.K. Counterfactual explanations for multivariate time series. Proceedings of the International Conference on Applied Artificial Intelligence; 2021 May 19–21; Halden, Norway; pp. 1-8. doi: 10.1109/ICAPAI49758.2021.9462056.