Trustworthy Machine Learning Operations for Predictive Maintenance Solutions

Kiavash Fathi[1,2], Tobias Kleinert[2], Hans Wernher van de Venn[1]

[1] *Institute of Mechatronic Systems, Zurich University of Applied Sciences, 8400 Winterthur, Switzerland*
fath@zhaw.ch, vhns@zhaw.ch

[2] *Chair of Information and Automation Systems for Process and Material Technology, RWTH Aachen University, 52064 Aachen, Germany*
kiavash.fathi@rwth-aachen.de, kleinert@plt.rwth-aachen.de

ABSTRACT

With the ever-growing capabilities of data acquisition and computational units in industry, development, and deployment of data-driven models (*e.g.*, predictive maintenance solutions) have become more abundant. However, if these models are not trained and maintained properly, they can be counterproductive as their predictions may be incorrect, unreliable, or difficult to interpret. In addition, unlike conventional software, the issues with such models often result in reduced productivity rather than traceable software errors. Therefore, we aim to use model performance evaluation measures introduced in trustworthy AI operations (TrustAIOps) to trigger re-evaluation of different parts of the data pipeline and the deployed data-driven model given machine learning operations (MLOps) requirements. We argue that by creating an ecosystem capable of monitoring different aspects of a data-driven solution by integrating and managing the implementation concepts in TrustAIOps and MLOps, it is possible to boost the performance of models given the constant changes induced by the specifications of Industry 4.0.

1. INTRODUCTION

Data acquisition and computational units improve daily which facilitate the development and deployment of data-driven approaches in Industry 4.0 settings. However, these data-driven models, when not trained and maintained properly, can be counterproductive as their predictions are not correct, reliable or interpretable. Unlike conventional software, the issues with model development manifest themselves in reduced productivity and not in other forms of traceable software error. In fact, when faced with during the run-time, they could be due to the errors from the data acquisition, data preprocessing, model training or model deployment submodels (Ashmore, Calinescu, & Paterson, 2021).

To ensure the acceptable performance of data-driven solutions, numerous implementation concepts have been introduced from the machine leaning operations (MLOps) society, which cover different aspects of preparing and deploying a data-driven solution. The following are some the most important characteristics of the models developed given MLOps requirements (Huyen, 2022):

1. **Reliability:** Correctness despite adversity
2. **Scalability:** Possibility of growth in complexity
3. **Adaptability:** Can cope with different data distribution shifts and business requirements
4. **Maintainability:** Documented and open to different tools

On the other hand, given the ever-growing application of machine learning solutions in different use cases, especially in safety-critical systems, performance criteria other than the accuracy have been promoted in research targeting trust worthy AI operations (TrustAIOps) which include but are not limited to (Li et al., 2023):

1. **Robustness:** Ability to deal with unseen data
2. **Generalization:** Distilling knowledge from limited training data for accurate predictions on unseen data
3. **Explainability:** Clarity on how a model makes decision
4. **Transparency:** Disclosing information about the model's lifecycle

As it can be seen in the above-mentioned characteristics, both MLOps and TrustAIOps put much emphasis on the performance of the deployed models given the possible changes in the data. These changes in Industry 4.0 settings are also relevant as there are many factors including production recipe, raw material vendor, product test unit fail/pass criteria, asset wear and tear, *etc.*, which can cause different types of data

distribution shifts. Predictive maintenance (PdM) as one of the important use cases of Industry 4.0 compliant solutions, is not an exception and requires tailored solutions for ensuring its effectiveness in an industrial setting.

1.1. Problem statement

How can the model performance evaluation measures introduced in TrustAIOps be used to trigger re-evaluation of different parts of the data pipeline and the deployed model given MLOps requirements? *(As a small remark; however, given the fact that the MLOps and TrustAIOps requirements cover numerous aspects of the PdM models, in the conducted studies, we consider only the characteristics listed above.)*

In the conducted research, we aim to introduce new implementation concepts which have proven to be useful for real industrial use cases in Europe and that are not properly addressed in the related work. In what follows pairs of MLOps and TrustAIOps, written as

$$TrustAIOps\ trigger \rightarrow MLOps\ requirement$$

are introduced with a specific implementation challenge for industrial PdM solutions:

1. **Robustness → Reliability:** Detecting previously unseen failures in the system
2. **Explainability → Scalability:** Interpretable model stacking
3. **Generalization → Adaptability:** Classifying different working conditions of an asset - Generation of run-to-failure data via simulation models
4. **Transparency → Maintainability:** Human-readable reports from different parts of the PdM solution

1.2. Research questions (RQs) and expected contributions

To elucidate further, given the complexity and high dimensionality of industrial data from different assets, how can

RQ 1. The model prediction certainty be correctly interpreted for out-of-training-distribution datapoints which represent previously unidentified failures of an asset? (see red blocks in Fig. 1). For inspecting the data-distribution shifts caused by changes in the working conditions refer to **RQ 3**.

RQ 2. The impact of different sources of uncertainty be minimized during model training using interpretable AI? (see grey blocks in Fig. 1)

RQ 3. Domain knowledge about different working conditions be included in data preprocessing and model training for enhanced data aggregation across different instances of the same production assets? (see green blocks in Fig. 1)

RQ 4. Lack of annotated data, *e.g.*, continuous data such as run-to-failure samples, be compensated using domain adaptation and simulation models? (see blue blocks in Fig. 1)

RQ 5. Human-readable reports be generated for increasing the transparency, *e.g.*, about how predictions are made and what data was used to train the model, of different submodels of the PdM solution, *esp.* for safety-critical system?

2. Conducted studies

In this section, a summary of the implemented solutions targeting parts of the first four **RQs**, specifically developed for the industry are presented. The solutions provided in this section adhere to the identical sequence as outlined in the **RQs**.

2.1. Detecting previously unidentified failures of an asset (Industry supported academic project)

It has been shown that the available data from different assets, even in case that they are abundant, normally do not cover different failure types that could occur in a system. Therefore, it is inevitable to monitor a PdM model in case data from a new working condition and/or failure type are exposed to it (Fig. 2). Despite numerous model calibration solutions, it has been observed that even models which are calibrated cannot demonstrate their certainty correctly when out-of-training-distribution data are fed into them. For PdM solutions, it is of utmost importance to inform the maintenance crew when a novel in the system has occurred as, otherwise, an exhaustive search is require for fault localization and diagnostics. In the conducted study, we have developed a post-hoc sample-based classification model built on top of the initial PdM solution that can detect previously unidentified failures in the system. The proposed method inspect the behavior of the PdM model, defined as the sequence of the PdM model certainty, and flags datapoints which indicate an anomaly in the PdM model behavior. The proposed method is tested on a demonstrator build by a company producing pneumatic components and has a mean accuracy of 94.35% (Fathi, Ristin, Sadurski, Kleinert, & van de Venn, 2024).

2.2. Reducing model uncertainty by interpretable model stacking (Industrial project)

Various changes in the production, *e.g.*, recipe updates, raw material vendor changes, improvements in quality test unit fail/pass criteria, *etc.*, impact the performance of the trained models given the potential data distribution shifts. If fact, with the adaptability in production as one of the main focuses of Industry 4.0, these changes reflect themselves in the data gathered from different assets which directly can impact the quality of the production. It is possible to counteract these changes in the gathered data by using different ensembling and model stacking techniques. In the conducted study we propose a novel approach for stacking the formerly trained

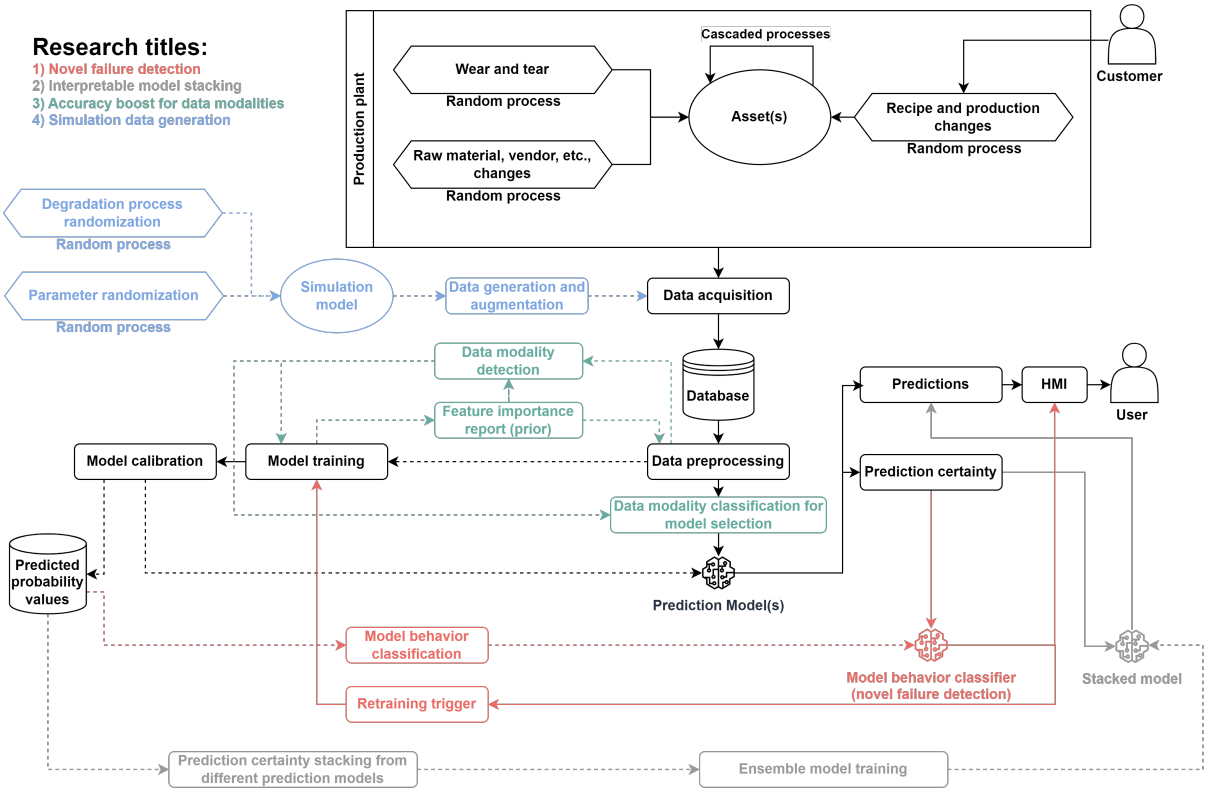

Figure 1. Overview of the proposed solution for TrustAIOps and MLOps integration in PdM

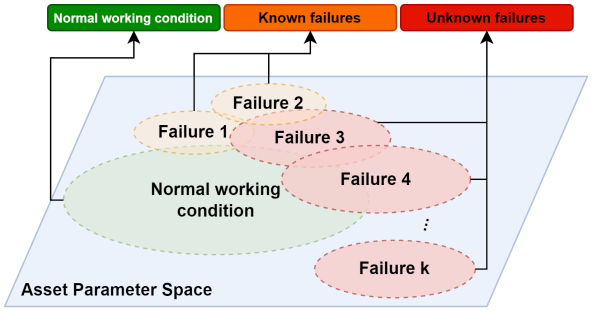

Figure 2. Asset parameter space and different known and unknown data modalities of the system

base learners. To avoid information loss due to prediction quantization of the base learners, in the proposed method we directly use the predicted probability values from the base learners and stack them using a linear regression model. The results demonstrate a 19.49% reduction in the binary estimated calibration error compared to convectional models which indicates the increased reliability of the final solution (Fathi, Stramaglia, et al., 2024).

2.3. Boosting model accuracy for different data modalities of an asset (Industrial project)

The constant changes in the production introduced in Subsection 2.2, can also lead to different dominant working condition of an asset which is also refereed to as data modality. In the conducted study, two instances of the same milling machine used for creating artificial bone joints of different sizes are examined to first detect and later to classify their different data modalities (see Fig. 3). Once different data modalities are distinguishable from one another, separate prediction models are trained for them which can increase the overall accuracy of the predictions up to 25.20%. In addition, for the data modality which forms the minority of the data from the asset, it is shown that by combining the corresponding data modalities from the two milling machines, it is possible to increase the accuracy for the aforementioned data modality up to 60.50%. In fact, by detecting corresponding data modalities, it is possible to address the problem of lack of annotated data for different instances of the same asset by simply sharing data from the same data modalities across the assets (Fathi, Sadurski, Kleinert, & van de Venn, 2023).

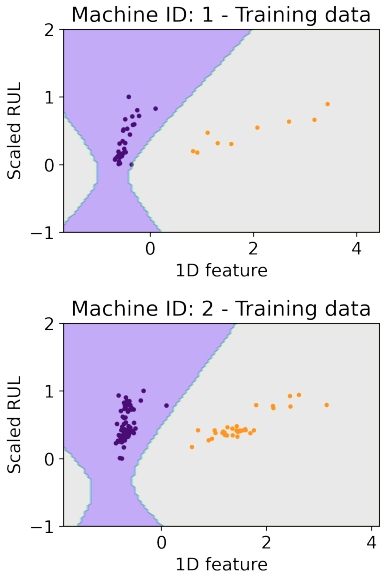

Figure 3. Decision boundaries of the trained model for classifying different data modalities of the asset

2.4. Data generation from simulation model for domain adaptation (Industrial project, paper under review)

Domain adaptation techniques developed for PdM normally focus on classification problem and neglect the regression problem of estimating the remaining useful life of an asset. In addition, they do not consider cases where the degradation of the asset is a random process itself either given the possibility of changes in the dominant failing component. Therefore, in the conducted study a novel approach for simulation data generation is introduced which is based on simulation parameter and data perturbation. It is shown how the proposed method can help cover different regions of the parameter space of the asset indicating different working conditions and parameterization of the asset (see Fig. 4). As a result, models trained with such data are more robust against signal reading manipulation and also demonstrate a more spread-out feature importance across a wider range of sensor readings while making predictions.

3. Future Work and Next Steps

Given the conducted studies listed above, it is inevitable to create an ecosystem which is capable of monitoring different aspects of a PdM solution by **integrating** and **managing** the implementation concepts introduced in Section 2. In fact, this ecosystem will use the introduced TrustAIOps concepts to ensure the expected performance of the PdM solution given MLOps requirements. One of the most important features of this ecosystem as introduced in **RQ 5** (see Section 1.2), is providing human-readable reports from different submodels of the PdM solution to ease its maintenance and debugging. One feasible solution for the aforementioned ecosystem is to create a metadata-based management system which is capable of tracking changes in different submodels of the deployed PdM solution. These changes are the essentially the response of the PdM solution for adapting to the new working conditions and/or previously unseen failures of the system. When done correctly, the proposed solution can be used as a foundation for data-driven PdM solutions of different assets including safety-critical systems.

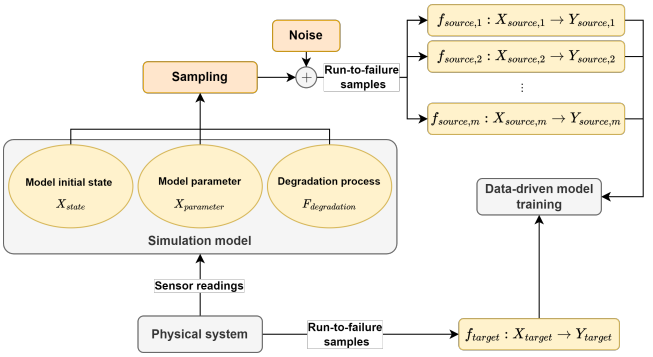

Figure 4. Domain adaptation via simulation parameter and data perturbation

References

Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, *54*(5), 1–39.

Fathi, K., Ristin, M., Sadurski, M., Kleinert, T., & van de Venn, H. W. (2024). Detection of novel asset failures in predictive maintenance using classifier certainty. *IEEE, 32nd Mediterranean Conference on Control and Automation (MED)*.

Fathi, K., Sadurski, M., Kleinert, T., & van de Venn, H. W. (2023). Source component shift detection classification for improved remaining useful life estimation in alarm-based predictive maintenance. *IEEE, 23rd International Conference on Control, Automation and Systems (ICCAS)*.

Fathi, K., Stramaglia, M., Ristin, M., Sadurski, M., Kleinert, T., Schönfelder, R., & van de Venn, H. W. (2024). Sustainability in semiconductor production via interpretable and reliable predictions. *IFAC, 12th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes*.

Huyen, C. (2022). *Designing machine learning systems*. "O'Reilly Media, Inc.".

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1–46.