Explainable MLOps: A Methodological Framework for the Development of Explainable AI in Practice

Annemarie Jutte[1,2]

[1] Saxion University of Applied Sciences, Enschede, The Netherlands
[2] University of Twente, Enschede, The Netherlands

Abstract

The field of explainable artificial intelligence (XAI) aims to increase the transparency of AI models by providing explanations for their reasoning processes. Valuable efforts have led to an increase in transparency. However, there are still blind spots in literature, specifically related to the use of XAI techniques in practice. To make the development of AI models truly explainable, transparency is required in each stage of the Machine Learning Operations (MLOps) workflow: data preparation, model development and model deployment. This research aims to mitigate issues in each stage, using case studies from the industry and health domains. The final objective is to provide an application-oriented methodological framework for the development of more transparent AI approaches.

Keywords

XAI, MLOps, data quality, application-oriented

1. Introduction

Artificial intelligence (AI) models, like those used for scheduling maintenance work [1] or diagnosing diseases [2], are taking on an increasingly significant role in critical decision-making processes. These models are often complex and it can be difficult to get insight into their internal workings [3]. The most complex of these models are referred to as 'black-box' models [3], where given an input, the decision of the model is known, but the reasoning behind it is unclear.

If model reasoning is not understood, the model may seem to perform well in a controlled environment, but we cannot be sure if it performs well for the right reasons. Consequently, we cannot be sure if it will perform well in the future. Due to the automation bias [4], users of said models may put (misguided) trust into them. Due to this trust, users may not notice unexpected behaviours. Efforts in the field of Explainable AI (XAI) have led to a wide range of available techniques to introduce explainability into AI models [5]. However, there are still blind spots within literature, specifically regarding the use of these techniques in practice.

Firstly, current XAI techniques generally do not take stakeholder needs into account. Instead, the focus tends to be on algorithmic possibilities [6]. Generally, machine learning engineers design techniques without considering other users [7], who may have limited to no knowledge of machine learning. As a consequence, the techniques are suitable to be used by machine

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

a.m.p.jutte@utwente.nl (A. Jutte)
0009-0004-9322-0674 (A. Jutte)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

learning engineers to get insight into their models during *model development*, but unsuitable for end-users during *model deployment*.

Secondly many issues within AI models come up due to issues within the data [8]. Following the 'garbage in, garbage out' principle, we cannot expect to develop high-quality models from low-quality data. To avoid investing resources into the training of models that are set up to fail, it would be valuable to mitigate data issues before training. Additionally, by bringing transparency into the data, stakeholders may recognize data issues without needing to understand complex models. Therefore, besides following the current trend in literature, where the focus is on making models transparent, this research will also consider transparency in *data preparation*.

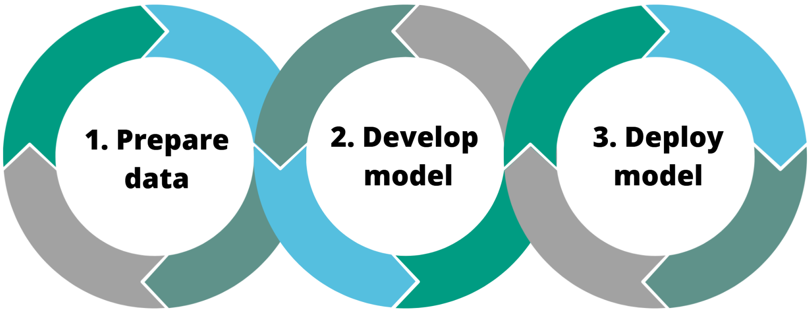

Figure 1: The MLOps workflow based on [9].

In this research, I argue, to make AI truly transparent, explainability should not be focused only on model development, but also on data preparation and model deployment. These stages relate to the MLOps (Machine Learning Operations) paradigm. MLOps leads to an iterative workflow, see Figure 1, based on the DevOps paradigm [10] in the software engineering field. The MLOps workflow consists of three stages: data preparation, model development and model deployment. MLOps aims to mitigate issues in the practical development and deployment of AI applications. It does so through rapid iterations where the different stages provide feedback to each other.

Currently, explainability is not explicitly part of the MLOps workflow. However, to develop truly transparent and robust AI models, it is imperative to be able to explain decisions in each MLOps stage. Therefore, this research aims to extend this workflow by bringing explainability into each step of the process:

- **Prepare data:** providing transparency into the data allows for verification of the data quality before training the model.
- **Develop model:** explanations allow machine learning engineers to detect issues with model reasoning to improve the model before passing the model on to domain-experts.
- **Deploy model:** when domain-experts interact with the model in critical decision-making processes, the model needs to justify the decisions it makes to prevent mistakes due to unexpected model behaviour and gain user trust.

The main objective of this research is to design a methodological framework that supports the development of end-to-end XAI approaches with an emphasis on application. This is a rather broad objective. Issues that may occur during practical AI development are in many aspects related to the modality of the data. Therefore, the choice is made to focus specifically on applications using sensor data. This choice is made since sensor data has been underrepresented in literature [11].

2. Related work

2.1. Explainable data preparation

The quality of the data is crucial for the quality of models trained using this data, low-quality data can introduce issues into AI models [8]. In literature on XAI, the focus is generally on explaining already developed AI models rather than first providing transparency for the data [5].

Literature that does directly address data issues, generally focuses on data issues related to biases towards certain outcomes. This focus is due to examples of AI discriminating against groups of people [12]. However, depending on the use case, there are also other issues related to data quality that have not been given much attention.

Calders and Žliobaitė [8] give three examples of ways in which data problems can present: biased sampling, incorrect labels and incomplete data. However, these issues can also be linked to different issues. A tired annotator may have given a subset of the data incorrect labels or the sensors in a dataset do not capture a phenomenon that is to be predicted. Black and van Nederpelt [13] present a list of 127 dimensions that determine data quality, including for example: consistency, precision, and reputation. This research will explore which dimensions are important for stakeholders and how to quantify data quality along these dimensions.

2.2. Explainable model development

In the model development stage, the data is used by machine learning engineers to develop AI models. XAI techniques can be used to validate whether the model is working as desired. Many techniques have been developed to uncover model reasoning [5].

Generally in development, a black-box model is initially developed with the aim of obtaining maximum predictive performance. Afterwards, a technique for post-hoc explainability [5] is added. However, there is often a trade-off between the predictive performance of the model and the level of explainability that can be reached [3]. More complex models often have a higher predictive performance, but are more difficult to interpret. Depending on the application, end users may accept some decrease in predictive performance, given the benefit of higher explainability.

Therefore, it could be beneficial to expand the current evaluation methods that focus on predictive performance, with factors such as explainability and data quality. Nauta et al. [14] presented evaluation methods for explainable models. Extending this evaluation with the evaluation of other model properties, like its computational cost and the quality of the used data, could give insight into the trade-off between model properties.

2.3. Explainable model deployment

During model deployment, the developed model is put into operation such that end-users can apply the model in their work. Most XAI approaches focus on extracting explanations for machine learning engineers, without making this information also understandable and informative for other users [7]. Therefore, many approaches are in fact not suitable for domain experts. Efforts have been made to uncover what different users need from explanations [6, 15] and techniques have been developed to provide more appropriate explanations [16, 17], but there are still steps to be made. Specifically, the explaining of temporal data to end users is underrepresented. Existing types of explanations are often not suitable for complex high-dimensional temporal data [18].

2.4. Explainable MLOps

The workflow in Figure 1 presents a general overview of the stages in MLOps. There are multiple models making the workflow more specific, for example CD4ML [19] or CRISP-ML(Q) [20]. They add stages such as business understanding and the development of front-end code. Depending on the findings in this research such a model will be adopted or adapted for the framework that will be the product of this research.

Previous research has started using such frameworks to expand the scope of XAI towards MLOps. For example, Dwivedi et al. [21] argue for both data explainability and model explainability in different phases. Kolyshkina and Simoff [22] presented a new workflow based on the interpretability requirements from different stakeholders for each stage. Tchuente et al. [23] present a six step approach to explainability.

The focus in these papers is mainly on organising existing research into different stages of the MLOps workflow. This research will instead work with use cases to tie the whole workflow together and make it more useful in practice. The research will include novel findings that mitigate existing gaps within each of the stages, as discussed previously in this section.

3. Research questions

The overarching purpose of this research is to develop a methodology for incorporating XAI into the MLOps workflow when working with sensor data. This leads to the following research question:

> *How can we design an application-oriented methodological framework that introduces explainability into the entire MLOps workflow for sensor data?*

In alignment with the respective stages in the MLOps workflow (see Figure 1), the following sub-questions are defined:

- RQ 1 - Explainable data preparation: How can issues related to data quality be quantified for more transparent decision-making?
- RQ 2 - Explainable model development: How to select AI models and explainability techniques that best match user requirements?

- RQ 3 - Explainable model deployment: How can AI models for sensor data be made explainable for domain experts?

4. Methodology

The research questions, discussed in the previous section, relate to four sub-studies, see Figure 2. Following the MLOps workflow, the overall process will be iterative in nature. Each study builds on the previous studies and new findings lead to updates of the previous studies.

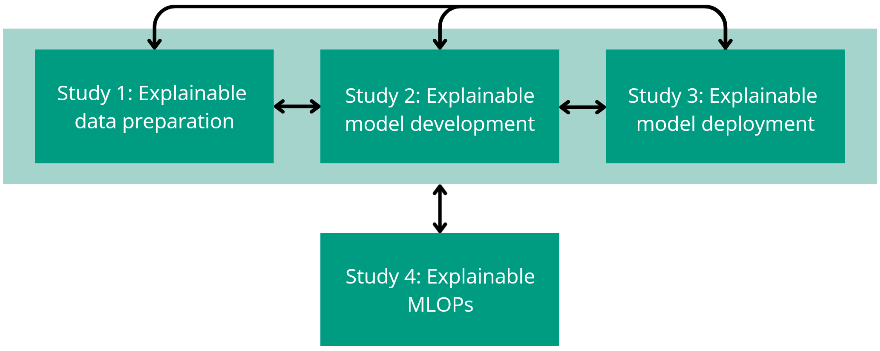
Study 1: Explainable data preparation
Study 2: Explainable model development
Study 3: Explainable model deployment
Study 4: Explainable MLOPs

Figure 2: Overview of the research structure.

This research will base itself around case studies. Drawing from the design science methodology [24], new artefacts (criteria, techniques or frameworks) will be designed based on findings from literature and findings from explorative user studies. The designed artefacts will be evaluated through user studies to evaluate their validity in a real world context. The goal is to generalize the findings into new theories for the state-of-the-art.

4.1. Case studies

The case studies will be used to uncover issues and desires from the application domain. For this research, the case studies will be from the industry and health domains. This research will focus on use cases these domains have in common: the prediction of events from sensor data.

In order to ensure case studies are suited for the purposes of this research, they will need to conform to the following criteria:

1. Data: sensor data must be available, the collection of this data is out of the scope of this project. Furthermore, data must be available from a variety of different sensors. The data should be sufficiently difficult to interpret, such that explainability is not inherent.
2. Technology readiness: an AI solution to the problem from the case study has been presented in at least one peer-reviewed paper, and this solution has been evaluated on a real-world dataset. The aim of this research is to design XAI methodologies, not to solve issues regarding predictive performance.

3. Infrastructure: in-house computational power is available or the company/institution is willing to run models in a secure cloud environment for deployment.

5. Limitations and threats

In this section we identify a number of threats to the success of this research, namely: data confidentiality, the maturity of evaluation techniques and bias in user studies.

Firstly, this research relies on data from real-world use cases. This type of data is generally subject to confidentiality, there might be restrictions to publishing research related to this data. A mitigation to this issue would be to use the confidential data for exploratory research, to uncover real-world problems, but replace the data with similar open data for publications.

Secondly, the focus of Study 2, see Section 4, is on selection criteria for AI models and XAI techniques. The goal is for these criteria to take multiple factors into account. However, the evaluation of explainability by itself is still in its early stages [14]. Therefore, the research may instead focus solely on evaluation methods for explainability.

Finally, the evaluation of this research relies on user studies. This type of evaluation is specifically sensitive to bias due to a small sample size. Therefore, the generalization of the results may be limited. To maximize the generalizability use cases from different domains are used, see Section 4.

6. Conclusion

In this research, I aim to design a novel application-oriented methodological framework for explainability in MLOps. To create this framework, each stage of the MLOps workflow will be researched in a study: explainable data preparation, explainable model development, and explainable model deployment. During these three studies, application-oriented issues will be explored in case studies and evaluated through user studies.

The first step in this research will be the study concerning explainable data preparation. This research will focus on the quantification of issues related to data quality by providing transparency into the data. The final step of this research will be to generalize the findings from the three studies into an application-oriented methodological framework.

The goal of this research is to support the development of transparent AI technologies in practice. The General Data Protection Regulation (GDPR) [25] states that citizens in the EU are entitled to a 'right to explanation' for algorithms that affect them.This research aims to contribute to the development data-driven algorithms that meet the requirements of this legislation.

Acknowledgments

I want to thank Faizan Ahmed, Jeroen Linssen and Maurice van Keulen for supervising this research. This publication is part of the project ZORRO with project number KICH1.ST02.21.003 of the research programme Key Enabling Technologies (KIC) which is partly financed by the Dutch Research Council (NWO).

References

[1] T. P. Carvalho, F. A. Soares, R. Vita, R. d. P. Francisco, J. P. Basto, S. G. Alcalá, A systematic literature review of machine learning methods applied to predictive maintenance, Computers & Industrial Engineering 137 (2019) 106024.

[2] S. M. D. A. C. Jayatilake, G. U. Ganegoda, Involvement of machine learning tools in healthcare decision making, Journal of healthcare engineering 2021 (2021).

[3] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[4] L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making?, International Journal of Human-Computer Studies 51 (1999) 991–1006.

[5] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems 263 (2023) 110273.

[6] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, Y. Li, Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle, in: Proceedings of the 2021 ACM Designing Interactive Systems Conference, 2021, pp. 1591–1602.

[7] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, P. Eckersley, Explainable machine learning in deployment, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 648–657.

[8] T. Calders, I. Žliobaitė, Why unbiased computational processes can lead to discriminative decision procedures, in: Discrimination and Privacy in the Information Society: Data mining and profiling in large databases, Springer, 2013, pp. 43–57.

[9] G. Symeonidis, E. Nerantzis, A. Kazakis, G. A. Papakostas, MLOps-definitions, tools and challenges, in: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2022, pp. 0453–0460.

[10] C. Ebert, G. Gallardo, J. Hernantes, N. Serrano, DevOps, IEEE software 33 (2016) 94–100.

[11] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable AI for time series classification: a review, taxonomy and research directions, IEEE Access 10 (2022) 100700–100724.

[12] J. Zou, L. Schiebinger, AI can be sexist and racist—it's time to make it fair, Nature 559 (2018) 324–326.

[13] A. Black, P. van Nederpelt, Dimensions of Data Quality (DDQ), 2020.

[14] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42.

[15] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) 103473.

[16] C. J. Cai, J. Jongejan, J. Holbrook, The effects of example-based explanations in a machine learning interface, in: Proceedings of the 24th international conference on intelligent user interfaces, 2019, pp. 258–262.

[17] V. Lai, Y. Zhang, C. Chen, Q. V. Liao, C. Tan, Selective explanations: Leveraging human input to align explainable ai, Proceedings of the ACM on Human-Computer Interaction 7

(2023) 1–35.

[18] E. Ates, B. Aksar, V. J. Leung, A. K. Coskun, Counterfactual explanations for multivariate time series, in: 2021 international conference on applied artificial intelligence (ICAPAI), IEEE, 2021, pp. 1–8.

[19] D. Sato, A. Wider, C. Windheuser, Continuous Delivery for Machine Learning, 2019. URL: https://martinfowler.com/articles/cd4ml.html, date accessed: 21-03-2024.

[20] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, K.-R. Müller, Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology, in: Machine learning and knowledge extraction, volume 3(2), MDPI, 2021, pp. 392–413.

[21] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 1–33.

[22] I. Kolyshkina, S. Simoff, Interpretability of machine learning solutions in industrial decision engineering, in: Australasian Conference on Data Mining, Springer, 2019, pp. 156–170.

[23] D. Tchuente, J. Lonlac, B. Kamsu-Foguem, A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications, Computers in Industry 155 (2024) 104044.

[24] A. R. Hevner, A three cycle view of design science research, Scandinavian journal of information systems 19 (2007) 4.

[25] European Parliament, Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.