Methods

Nadia Burkart*, Danilo Brajovic and Marco F. Huber

Explainable AI: introducing trust and comprehensibility to AI engineering

Erklärbare KI: Einführung von Vertrauen und Nachvollziehbarkeit in das KI-Engineering

https://doi.org/10.1515/auto-2022-0013
Received February 7, 2022; accepted July 21, 2022

Abstract: Machine learning (ML) rapidly gains increasing interest due to the continuous improvements in performance. ML is used in many different applications to support human users. The representational power of ML models allows solving difficult tasks, while making them impossible to be understood by humans. This provides room for possible errors and limits the full potential of ML, as it cannot be applied in critical environments. In this paper, we propose employing Explainable AI (xAI) for both model and data set refinement, in order to introduce trust and comprehensibility. Model refinement utilizes xAI for providing insights to inner workings of an ML model, for identifying limitations and for deriving potential improvements. Similarly, xAI is used in data set refinement to detect and resolve problems of the training data.

Keywords: explainable AI, machine learning, model refinement, data set refinement

Zusammenfassung: Maschinelles Lernen (ML) gewinnt aufgrund kontinuierlicher Leistungssteigerungen zunehmend an Interesse. ML wird in vielen verschiedenen Anwendungen eingesetzt, um menschliche Nutzer zu unterstützen. Die Repräsentationsmächtigkeit von ML-Modellen ermöglicht es, schwierige Aufgaben zu lösen, macht es aber unmöglich, dass die resultierenden Modelle von Menschen verstanden werden. Dies bietet Raum für mögliche Fehler und schränkt das volle Potenzial von ML ein, da ein Einsatz in kritischen Umgebungen nicht möglich ist. In dieser Arbeit schlagen wir vor, erklärbare KI (xAI) sowohl für die Modell- als auch für die Datensatzverfeinerung einzusetzen, um Vertrauen und Verständlichkeit zu schaffen. Bei der Modellverfeinerung wird xAI eingesetzt, um Einblicke in das Innenleben eines ML-Modells zu erhalten, um Einschränkungen zu erkennen und um potenzielle Verbesserungen abzuleiten. Ebenso wird xAI bei der Datensatzverfeinerung eingesetzt, um Probleme mit den Trainingsdaten zu erkennen und zu beheben.

Schlagwörter: Erklärbare KI, maschinelles Lernen, Modellverfeinerung, Datenverfeinerung

1 Introduction

In many engineering fields, artificial intelligence (AI) and its sub-field machine learning (ML) play a major role in advancing the functionality of products and processes. For instance, achieving automated driving at SAE[1] level 4 or 5 is considered unrealistic without methods from ML [8]. The same holds for Industry 4.0 techniques like human-robot-collaboration or predictive maintenance. At the same time, the utilization of ML in such domains is often seen skeptical as the most accurate ML models like deep neural networks or random forests behave like a *black box* to the human, i. e., the internal decision processes and calculations are opaque [2]. This makes it demanding or even impossible for humans to understand or proof the correct functionality of an AI system.

A common solution used in engineering, especially for safety critical applications like automated driving, is to use verification based on data-driven or formal methods [1]. Data-driven approaches provide statistical statements about the correct functionality, while formal methods mathematically prove or disprove the correctness [11]. However, the definition of the properties to be verified and the search for the cause of errors in the event of a violation of these properties are challenging [1]. It further becomes

***Corresponding author: Nadia Burkart,** Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany, e-mail: nadia.burkart@iosb.fraunhofer.de
Danilo Brajovic, Marco F. Huber, Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Department Cyber Cognitive Intelligence (CCI), Stuttgart, Germany, e-mails: danilo.brajovic@ipa.fraunhofer.de, marco.huber@ipa.fraunhofer.de

1 The SAE (Society of Automotive Engineers) J3016 standard describes the classification and definition of terms for motor vehicles with automated driving systems known as SAE levels.

only limited evident to what extent wrong system behavior is originating from systematic errors in the data set being used to train an ML model or due to the learned model itself.

The use of explainable AI (xAI) is intended to alleviate these issues. xAI methods can be used to draw conclusions about the learned system behavior and help to build up a better understanding of the system, which resolves the black box problem and increases trust into the system. The insights derived from xAI methods can be beneficial for the definition of a formal verification process as well as for the optimization of the AI system design [15]. In this paper, we give an introduction on how xAI methods allow refining both the ML model and the data set for optimizing an AI system. For this purpose, in Sec. 2 an overview of xAI is provided. Model and data set refinement are addressed in Sec. 3. The paper closes with conclusions and an outlook to future work.

2 Explainable AI

ML algorithms usually construct complex models to solve specific problems and have become very accurate over the past few years. The downside of the complex models are the lack of ability to explain the reasoning, which is oftentimes based on millions or even billions of parameters and calculations. In the past, the performance measure of interest was only a high accuracy. In the literature, there are many examples where the model has a high accuracy, but the decisions are based on wrong correlations [14]. An unnoticed aspect often was how these models work and what the reasons behind their decisions were. Without this information there is plenty of room for errors that can cause serious harms. While the performance of such systems is beneficial to a wide variety of use cases, the true potential is limited. For the application in, e. g., predictive maintenance scenarios in industrial manufacturing reliable models are needed. If an ML model is trained to support a maintenance team, it is important to know the reasons why a system could fail in order to take the right measures. Therefore, it is mandatory for the maintainer of the production machine to know which parameters exactly influence the failure and how the failure can be avoided at an early stage.

To achieve this, decisions made by ML models need to be explainable. The research field of Explainable Artificial Intelligence (xAI) attempts to make specific decisions of the model or the entire model explainable for different stakeholders, e. g., developers, domain experts, end users, etc. With xAI it is possible to evaluate the quality of an ML model and its predictions. Understanding these models allow developers to create better models and enables users to decide whether to trust an AI system.

There were few early works, such as the feature relevance of random forests [3] that are considered to be a milestone in this area [13]. Over time, it became increasingly apparent that the explainability of the models is crucial for the success of the applications in different fields. Burkart and Huber [2] introduced a categorization of different xAI approaches with an extensive literature review of current approaches. xAI approaches can be categorized by ante-hoc and post-hoc. Ante-hoc means that the trained model itself is either interpretable by nature or optimized for explainability. The resulting model is also often called a *white box model*, e. g., a small decision tree. Post-hoc approaches are applied on top of the black box to generate explainability for the entire model or specific instances of the data.

Moreover, the approaches can be categorized into model-specific and model-agnostic. Model-specific approaches use the internals of the model to generate explanations. Ante-hoc approaches are always specific. Model-agnostic approaches just work with the input data and the output data of the black-box model. Further, the approaches can be divided into global and local. Ante-hoc approaches are always global because the entire model is in focus. Global post-hoc approaches usually create a surrogate model for an already existing black box, which is itself globally explainable, but behaves exactly like the black box. Therefore the surrogate model can provide explanations for the behaviour of the black box. The important factors here are not only accuracy and explainability, but also the surrogate fidelity. It is very crucial, that the surrogate model resembles the black box model as close as possible. There are also global approaches that can directly infer explanations from the black box, e. g., a random forest feature importance [3].

3 Model and data set refinement

To build a machine learning model it is best practice to decompose the engineering process into various stages commencing from the use case definition stage and ending with the deployment of the final AI system. This so-called *machine learning pipeline* is depicted in Fig. 1. Similar decompositions can be found in process models like CRISP-DM [18], KDD [5] or PAISE [7]. It can be easily seen that the majority of the stages of this pipeline are concerned with

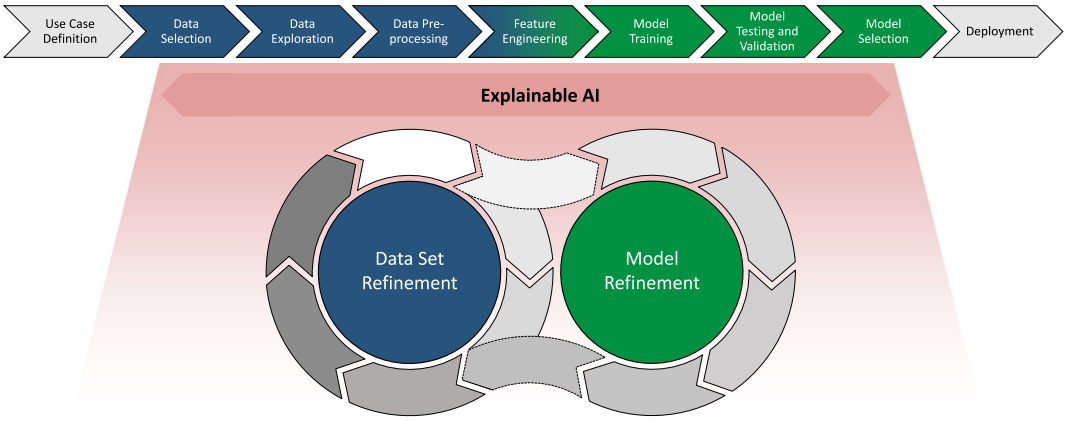

Figure 1: Model and data set refinement utilizing xAI along the various stages of the machine learning pipeline.

the data and the model. Research in machine learning often focuses on the model stages only, e. g., by proposing improved learning or hyperparameter optimization techniques. A strong focus on the model can also be observed in xAI as became apparent in Sec. 2. However, just recently a new paradigm named *data-centric machine learning* was proposed by Andrew Ng:[2]

> "Instead of focusing on the code, companies should focus on developing systematic engineering practices for improving data in ways that are reliable, efficient, and systematic. In other words, companies need to move from a model-centric approach to a data-centric approach."

Thus, the key idea of the paradigm is to systematically improve the data sets to increase the performance of the machine learning model. Hence, we propose to extend the current model-centric focus of xAI also on the data sets. In doing so, xAI techniques are exploited in order to gain insights, understand limitations and provide actionable recommendations to data scientist and AI engineers for refining both data and models during the development of an AI system.

In the following, we describe how xAI can be used for model refinement and data set refinement. We further highlight benefits of the combined approach and its current limitations.

3.1 Model refinement

There are different approaches to introduce explainability within the model refinement stage that help, e. g., data scientists or other stakeholders to improve the model. Most of the current xAI approaches are used within the model testing and validation phase (see Fig. 1). To make the use of xAI for model refinement understandable, we will give an example from the predictive maintenance domain. Here, a random forest model is trained on a synthetic data set [12] to predict a machine failure. The data set consists of 10.000 data instances with 14 features (torque, air temperature, etc.).

In general, xAI approaches generate different outputs that the stakeholder can interpret to improve the model. Exemplary outputs are linear models, decision trees, rule sets, or attribution maps. A linear model assigns a weighting factor to each input feature. A rule set comprises of `IF condition THEN label ELSE different label` statements. Tree-based models build a decision tree from the input features. Attribution maps highlight important parts of the input and is usually applied to image or text data.

One of the most fundamental questions that can help improve the ML model is which features have the greatest influence on the prediction. This concept is known as *feature importance* and can be extracted for instance by means of a linear model or a tree-based model. Features can be relevant to a prediction on a local and a global scope.

Local post-hoc approaches focus on generating feature importance scores for single predictions. The aim is to make the decision-process behind a single prediction comprehensible, without explaining the entire model. One of the most famous frameworks for local post-hoc explanations is LIME [14]. LIME is a model-agnostic approach that generates feature importance for any black-box model by training a local surrogate model for a specific prediction. The stakeholder thereby can review the decision process for specific instances.

2 https://landing.ai/data-centric-ai, last viewed January 22, 2022.

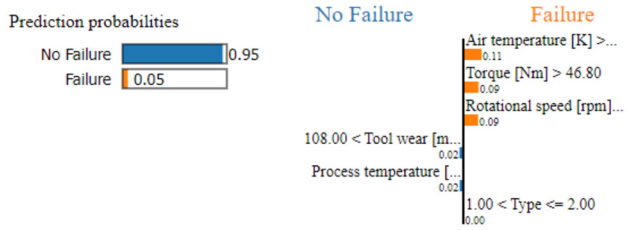

Figure 2: LIME explanation for a test instance.

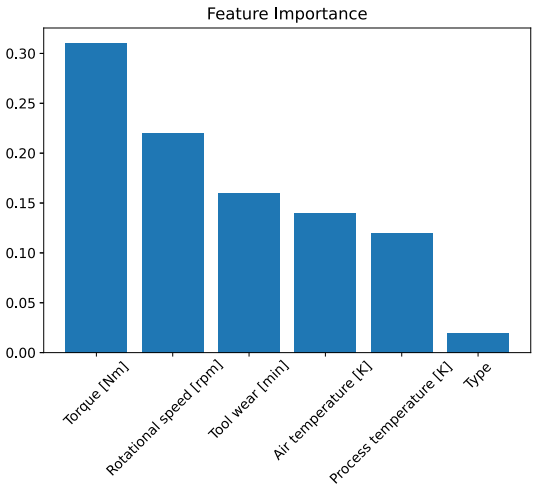

Figure 3: Random forest feature importance.

Fig. 2 illustrates the output of LIME. The model predicts with 95 % probability that no failure will happen and with 5 % probability that a failure will occur. Now LIME supports the stakeholder to understand what features will probably influence the failure case (coloured in orange). Using this information, the stakeholder can determine whether the model matches, for example, known failure cases and expert knowledge about the production process.

Global approaches in contrast infer an explanation for the entire model. A commonly used global approach is the random forest feature importance [3]. A random forest consists of a set of decision trees that are jointly used to perform a prediction. The random forest feature importance relies on the information gain and Shannon entropy metric, respectively, which is used to selected the splitting features during learning the various decision trees. This metric can be accumulated per feature over all trees and thus, indicates the most important features the entire random forest is relying on. Fig. 3 shows the random forest feature importance for the predictive maintenance use case. For example, if a domain expert realizes that the model assigns a high importance score to a feature that is unlikely to indicate a failure, the feature could be removed to improve the model.

3.2 Data set refinement

Often, it is not the model choice that leads to poor performance of a machine learning component but the data set itself. Recent work actually suggests that many properties that a model learns from a data set are model agnostic and, hence, a property of the data set rather than the model [17]. The corollary is that insights emerging from understanding the model with the introduced methods of xAI may as well be used to improve the data set itself. For computer vision use cases, one of the easiest ways to do so is to apply the gradient to the input image in order to maximize a certain output class [16]. If one is maximizing the prediction for the class *grass* and the image starts showing sheep this is an indicator that the training data contains too much images of grass with sheep. Hence, the data set has to be extended with more images of grass without sheep. A related concept are counterfactuals. For a counterfactual, the question is how the input has to be altered in order to change the models prediction. In the previous example, this could mean that the grass in the background has to be removed so that the models output changes from sheep to something else. As before, such an explanation indicates that the training data contains too much examples of sheep on grass. However, here it is also possible to add the counterfactual example of a sheep in front of another background to the training data set.

The above method describes how to detect and resolve concrete errors in a data set once they are found and only when they are understandable to a human. Unfortunately, this is not always the case. Instead of showing sheep, the maximized input image in the previous example could just contain noise. In this case, it is unclear whether the model is good or how to improve it. The question how a good and representative data set for a task actually looks like is still open. One interesting concept related to this are the long tails of many data sets that contain atypical and rare samples [9, 4]. A sample may be interpreted as rare or singleton if it is the only one showing a certain concept. Usually, such a sample can be detected because it is miss-classified as soon as it is removed from the training data. A data set containing many such samples can be considered as undesirable because it is vulnerable to privacy attacks [4] and because the absence of a single image can significantly degrade the performance. Future research should address how to estimate such samples effectively and how to extend the data set meaningfully.

Finally, a set of recent methods can be applied to remove many unimportant samples from a data set without hurting performance [10, 6, 19, 17]. This is interesting because it works between different neural network ar-

chitectures. Hence, unimportant samples can first be removed with the help of a smaller model and then, a larger model can later be trained on the reduced data set. These methods can be interpreted as shifting explainability of the model to explainability of a data set. One reason for this view is that the Shapley values known from model explainability are applied to the data [6].

4 Conclusion and future work

In order to engineer AI systems that maintain a high level of performance but also allow different stakeholders to understand, appropriately trust, and effectively manage AI systems, we consider it mandatory to not only focus on the ML model itself. Of at least the same importance is to understand the influence of the data on the ML model. For this purpose, we proposed to employ techniques from xAI on all stages of the ML pipeline to allow for model and data set refinement. Our future work within the field of model and data set refinement is concerned with the developing an engineering tool suite that allows easily incorporating various xAI methods into the ML pipeline development. Further, we will conduct user studies to showcase the benefit of this tool suite.

Funding: This work was supported by the Baden-Wuerttemberg Ministry for Economic Affairs, Labour and Tourism (Projects KI-Fortschrittszentrum "Lernende Systeme und Kognitive Robotik" and Competence Center KI-Engineering CC-KING).

References

1. Burton, S. and R. Hawkins. 2020. Assuring the safety of highly automated driving: State-of-the-art and research perspectives. Technical report, University of York.
2. Burkart, N. and M. Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research (JAIR)* 70: 245–317.
3. Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5–32.
4. Feldman, V. 2020. Does learning require memorization? A short tale about a long tail. In: *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 954–959.
5. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth. 1996. From data mining to knowledge discovery in databases. *AI Magazine* 17(3): 37.
6. Ghorbani, A. and J. Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In: *36th International Conference on Machine Learning, ICML 2019*, 2019 June, pp. 4053–4065.
7. Hasterok, C., J. Stompe, J. Pfrommer, T. Usländer, J. Ziehn, S. Reiter, M. Weber and PAISE Till Riedel. 2021. Das Vorgehensmodell für KI-Engineering. White paper, Kompetenzzentrum KI-Engineering CC-KING.
8. Huval, B., T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, F.A. Mujica, A. Coates and A. Ng. 2015. An empirical evaluation of deep learning on highway driving. arXiv:1504.01716.
9. Jiang, Z., C. Zhang, K. Talwar and M.C. Mozer. 2020. Characterizing structural regularities of labeled data in overparameterized models.
10. Koh P.W. and P. Liang. 2017. Understanding black-box predictions via influence functions. In: *34th International Conference on Machine Learning, ICML 2017*, pp. 2976–2987.
11. Liu, C., T. Arnon, C. Lazarus, C. Strong, C. Barrett and M.J. Kochenderfer. 2021. Algorithms for verifying deep neural networks. *Foundations and Trends® in Optimization* 4(3–4): 244–404.
12. Matzka, S. 2020. Ai4i 2020 predictive maintenance dataset. UCI Machine Learning Repository.
13. Molnar, Ch. 2020. Interpretable machine learning. Lulu.com.
14. Tulio Ribeiro, M., S. Singh and C. Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.
15. Salay, R. and K. Czarnecki. 2018. Using machine learning safely in automotive software: An assessment and adaption of software process requirements in ISO 26262. arXiv:1808.01614.
16. Simonyan, K., A. Vedaldi and A. Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: *Workshop at International Conference on Learning Representations*.
17. Toneva, M., A. Sordoni, R. Tachet des Combes, A. Trischler, Y. Bengio and G.J. Gordon. 2018. An empirical study of example forgetting during deep neural network learning, pp. 1–19. Published in ICLR 2019. Arxiv: https://arxiv.org/abs/1812.05159.
18. Wirth, R. and J. Hipp. 2000. CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39.
19. Yoon, J., S. Arik and T. Pfister. 2020. Data valuation using reinforcement learning. In: *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119. pp. 10842–10851. Arxiv: https://arxiv.org/abs/1909.11671.

Bionotes

Nadia Burkart
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany
nadia.burkart@iosb.fraunhofer.de

Nadia Burkart received her Bachelor degree (2011) and Master degree (2013) in business informatics from the University of Applied Science in Karlsruhe. 2013 she started as a research scientist at the Fraunhofer IOSB in Karlsruhe in the field of decision support systems. Since 2021 she is leading the research group Applied Explainable AI at Fraunhofer IOSB. In this context she is working on various projects on explainable machine learning solutions in several domains. Besides her main project business she finished her PhD thesis in the field of explainable machine learning in 2021.

Danilo Brajovic
Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Department Cyber Cognitive Intelligence (CCI), Stuttgart, Germany
danilo.brajovic@ipa.fraunhofer.de

Danilo Brajovic received a Bachelor's degree in computer science and both a Master's degree in cognitive and in computer science from Tübingen University in 2017, 2020 and 2021. Currently, he is working in the Center for Cyber Cognitive Intelligence (CCI) at the Fraunhofer IPA in Stuttgart, Germany. His research is focused around safe AI in industrial applications.

Marco F. Huber
Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Department Cyber Cognitive Intelligence (CCI), Stuttgart, Germany
marco.huber@ipa.fraunhofer.de

Marco Huber received his diploma, Ph. D., and habilitation degrees in computer science from the Karlsruhe Institute of Technology (KIT), Germany, in 2006, 2009, and 2015, respectively. From June 2009 to May 2011, he was leading the research group Variable Image Acquisition and Processing of the Fraunhofer IOSB, Karlsruhe, Germany. Subsequently, he was Senior Researcher with AGT International, Darmstadt, Germany, until March 2015. From April 2015 to September 2018, he was responsible for product development and data science services of the Katana division at USU Software AG, Karlsruhe, Germany. At the same time he was adjunct professor of computer science with the KIT. Since October 2018 he is full professor with the University of Stuttgart. He further is director of the Center for Cyber Cognitive Intelligence (CCI) and of the Department for Image and Signal Processing with Fraunhofer IPA in Stuttgart, Germany. His research interests include machine learning, planning and decision making, image processing, data analytics, and robotics. He has authored or co-authored more than 100 publications in various high-ranking journals, books, and conferences, and holds two U.S. patents and one EU patent.