machine learning & knowledge extraction

MDPI

Editorial

Explainable Machine Learning

Jochen Garcke [1,2,†] and Ribana Roscher [3,4,*,†]

[1] Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany
[2] Fraunhofer Center for Machine Learning and Fraunhofer SCAI, 53757 Sankt Augustin, Germany
[3] Institute of Geodesy and Geoinformation, University of Bonn, 53113 Bonn, Germany
[4] Institute of Bio- and Geosciences, Forschungszentrum Jülich GmbH, 52428 Jülich, Germany
* Correspondence: ribana.roscher@uni-bonn.de
† These authors contributed equally to this work.

Machine learning methods are widely used in commercial applications and in many scientific areas. There is an increasing demand to understand the way a specific model operates and the underlying reasons for the decisions produced by the machine learning model. Increasing trust and justifying what has been learned; enhancing control; improving what has been learned by, for example, robustifying the model; and discovering novel insights and gaining new knowledge are a few of many reasons for seeking explanations [1,2]. In the natural sciences, where ML is increasingly employed to optimize and produce scientific outcomes, explainability can be seen as a prerequisite to ensure the scientific value of the outcome. In societal contexts, the reasons for a decision often matter. Typical examples are (semi-)automatic loan applications, hiring decisions, or risk assessment for insurance applicants. Here, in addition to regulatory reasons and fair decision making, one wants to gain insight into why a model gives a certain prediction and how this relates to the individual under consideration. For engineering applications, where ML models are deployed for decision-support and automation in potentially changing environments, an assumption is that with explainable ML approaches, robustness and reliability can be realized more easily. While machine learning is employed in numerous projects and publications today, the vast majority of applications are not concerned with aspects of interpretability or explainability. This Special Issue aims to present new approaches to explainable ML and to show the potentials of the methods in different disciplines. This Special Issue includes five new contributions with topics from different disciplines: the use of different explainable machine learning approaches to analyze an air quality benchmark dataset [3], the interpretation and explanation of object detection in natural images [4], the introduction of an interpretable approach for residual neural networks based on an implicit architecture [5], the identification of concepts in images to describe the only vaguely defined class 'landscape scenicness' [6], and the systematic incorporation of domain knowledge in a hybrid model to estimate soil moisture [7]. The works not only address different applications but also use a variety of explainable machine learning tools to achieve their goals. The works not only address different applications but also use a variety of explainable machine learning tools to achieve their goals. In [3,4], post hoc, model-agnostic approaches such as LIME and SHAP were used, while [3] also used model-specific techniques such as neural network weight analysis and random forests. Ref. [6] learned concept activation vectors from known datasets to test to what extent they can be found in new data. Breen et al. [7] focused on integrating expert knowledge from a specific domain to make the model more understandable and robust.

From these contributions, it is clear that explainable machine learning can lead to further insights that go beyond the estimation itself. The insights concern the model and its functioning and decision process [3–5], the reliability and robustness of the obtained results [7], and the input data such as the suitability as a basis for learning an estimation model [3] or novel characteristics and relations that enhance the understanding of single instances [6]. We can currently see a rapid development of new explainable machine

check for updates

Citation: Garcke, J.; Roscher, R. Explainable Machine Learning. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 169–170. https://doi.org/10.3390/make5010010

Received: 23 December 2022
Revised: 10 January 2023
Accepted: 11 January 2023
Published: 17 January 2023

learning approaches. However, their suitability, especially for scientific data, is still poorly explored and their thorough evaluation remains an open research question.

Author Contributions: Writing –Original Draft Preparation, J.G. and R.R.; Writing– Review & Editing, J.G. and R.R. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The guest editors would like to thank all the authors for their excellent contributions, the reviewers for their constructive comments and the editors of machine learning and knowledge extraction for their kind help.

Conflicts of Interest: The authors declare that there are no conflict of interest.

References

1. Holzinger, A. The next frontier: AI we can really trust. In Proceedings of the Joint European Conference On Machine Learning Furthermore, Knowledge Discovery in Databases, Bilbao, Spain, 13–17 September 2021; pp. 427–440.
2. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160.
3. Stadtler, S.; Betancourt, C.; Roscher, R. Explainable Machine Learning Reveals Capabilities, Redundancy, and Limitations of a Geospatial Air Quality Benchmark Dataset. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 150–171. [CrossRef]
4. Sejr, J.; Schneider-Kamp, P.; Ayoub, N. Surrogate Object Detection Explainer (SODEx) with YOLOv4 and LIME. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 662–671. [CrossRef]
5. Reshniak, V.; Webster, C. Robust Learning with Implicit Residual Networks. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 34–55. [CrossRef]
6. Arendsen, P.; Marcos, D.; Tuia, D. Concept Discovery for The Interpretation of Landscape Scenicness. *Mach. Learn. Knowl. Extr.* **2020**, *2*, 397–413. [CrossRef]
7. Breen, K.; James, S.; White, J.; Allen, P.; Arnold, J. A Hybrid Artificial Neural Network to Estimate Soil Moisture Using SWAT+ and SMAP Data. *Mach. Learn. Knowl. Extr.* **2020**, *2*, 283–306. [CrossRef]