Research Article

Andreas Holzinger*

Explainable AI and Multi-Modal Causability in Medicine

https://doi.org/10.1515/icom-2020-0024

Abstract: Progress in statistical machine learning made AI in medicine successful, in certain classification tasks even beyond human level performance. Nevertheless, correlation is not causation and successful models are often complex "black-boxes", which make it hard to understand *why* a result has been achieved. The explainable AI (xAI) community develops methods, e. g. to highlight which input parameters are relevant for a result; however, in the medical domain there is a need for causability: In the same way that usability encompasses measurements for the quality of use, causability encompasses measurements for the quality of explanations produced by xAI. The key for future human-AI interfaces is to map explainability with causability and to allow a domain expert to ask questions to understand why an AI came up with a result, and also to ask "what-if" questions (counterfactuals) to gain insight into the underlying *independent* explanatory factors of a result. A multi-modal causability is important in the medical domain because often different modalities contribute to a result.

Keywords: explainable AI, Human-Centered AI, Human-AI interfaces

1 Introduction: Deep Learning Success Examples in Medical AI

To reach human-level AI is the quest of AI researchers since the emergence of this field [1]. Research in the last decades has proved it to be very difficult and the progress has been slow, despite great success in statistical machine learning theory [2] and statistical learning practice [3]. Recently, there have been many practical successes of deep neuronal network learning [4] due to the availability of large amounts of training data sets and computational power. In the health domain there are many different areas, where AI can help, e. g. in diagnostics and decision making, drug discovery, therapy planning, patient monitoring, risk management, areas dealing with "big data" such as the analysis of *omics data, including genomics, proteomics, metabolomics, and many others [5]. One particular relevant field is medical image analysis, including AI-applications in pathology [6], radiology [7], dermatology [8], ophthalmology [9], oncology [10], and many other medical application fields.

Let us look at a meanwhile classic work, presented in 2017 by the group of Sebastian Thrun from Stanford, which was sold under "Beyond human-level performance" [11] and was popularized in the news in Europe as "AI is better than doctors". What did they do? They classified skin lesions using a single convolutional neural network (CNN), trained it end-to-end from the derma images directly, and used only pixels and disease labels as inputs. For pretraining they used 1.3 million images from the 2014 ImageNet challenge. Then, they used 130 thousand clinical images consisting of approximately 2000 different diseases, reaching 92 % average classification performance, on par with human dermatologists or even better. If we consider that the algorithm does not get tired this is really an amazing result, considered by medical doctors as a very good performance. However good these results may be, pressing questions are raised: *"Why can AI solve some tasks better than humans?"*, *"Why does the AI achieve such results?"*, *"Which underlying factors are contributing to the result?"*, or "*What if I change, replace, disturb, remove input data?"*, or more technically: *"What if the input data changes counterfactually …?"* This needs to consider and examine desired properties of methods, including *fidelity* [12, 13], *interpretability* [14], *parsimony* [15], and *generalizability* [16].

A very recent work from the Princess Margaret Cancer Center in Toronto in the field of histopathology [17] goes one step in this direction: They also applied a CNN to a repository of 840 thousand histopathological image tiles and learned representations into a 512-dimensional feature vector. The novelty here is that they showed that machine-generated features correlate with certain morphological constructs and ontological relationships gen-

***Corresponding author: Andreas Holzinger,** Human-Centered AI Lab, Institute for Medical Informatics & Statistics, Medical University Graz, Graz, Austria; and xAI Lab, Alberta Machine Intelligence Institute, Edmonton, Canada, e-mail: andreas.holzinger@medunigraz.at, ORCID: https://orcid.org/0000-0002-6786-5194

erated by humans. Why is this important for us? Because highlighting such overlaps between human thinking and machine "thinking" can contribute to what are currently top issues in the machine learning community: i) to eliminate bias and to improve algorithms robustness, and ii) to make the results retraceable, hence explainable in order to meet the quest of accountability of medical AI.

Despite all these successes, one of the most pressing problems is in robustness, i. e. in overcoming the "brittleness" of current AI systems, because true human-level AI requires computational approaches that are able to deal with "common sense" situations [18] and to "think" and "act" like humans. Many advances have resulted from using deep neural networks trained end-to-end in such tasks. Despite their biological inspiration and the impressive results mentioned before, these systems differ from human intelligence enormously. Besides lacking robustness and generalization, current approaches are unable to build causal models in order to support deep understanding [19]. Consequently, to make such approaches even more successful we need further work to make them *robust* [20], [21], *understandable* [22], and *interpretable* for a human expert [14]. The aim is to take advantage of the respective benefits of both statistical machine learning methods and model-based approaches, or more precisely: The aim is to integrate existing a-priori knowledge and human experience into statistical learning methods, thereby combining them synergistically in a hybrid approach to exploit the full benefits of data-driven methods without ignoring already acquired knowledge an human expertise. Here, a human-in-the-loop can be (sometimes, of course not always) helpful as we will discuss in section 3. Before that, we briefly discuss some basics of explainability and causability.

2 Explainability and Causability

The field of explainable AI is meanwhile very popular [23], [24], [25], [26], and the explainable AI (xAI) community is very active in developing various methods to help making such "black box" approaches, as outlined in the introduction, retraceable, understandable, and human interpretable.

It is important to note that results are interpretable when they classify objects on the basis of features that *a human can understand* [27]. Current approaches to explaining the decisions of deep learning for medical tasks have focused on visualising the elements that have contributed to each decision, which can be done e. g. via interactive heatmaps [28], [29]. Such "mechanical explanations" to highlight which input is relevant to an obtained output can be reached by using various methods: The simplest method works with gradients as multi-variable generalization of the derivative, where the neural network is seen as a function and the explanation relies on the function's gradient, which is available from the backpropagation algorithm [30]. Another possibility is to use decomposition methods (luckily our world is compositional), e. g. pixel-wise relevance propagation [31], layer-wise relevance propagation [32], or deep Taylor decomposition [33], which also works on graph-based data [34]. Other methods include deconvolution by reversing the effects of convolution and bringing out from two functions a third function which is then the product of both, guided backpropagation, and the use of so-called concept activation vectors [35], [36], [37], [38].

All these methods are excellent pre-processing steps, however, in a way that a medical expert can understand the *causality* of a learned representation and use it for *medical decision support* the xAI methods need to be developed even further. Let us note that xAI (or "explainability") deals with the implementation of methods to enable retraceability, transparency, and interpretability of so-called "black-box" methodologies. The currently best performing methods, as we have seen in the best-practice examples in the introduction above, are of such kind. Unfortunately, it is not an option just to say *"stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead"* as stated by Cynthia Rudin [14], because this would mean not to use the currently best performing methods.

However, in the biomedical domain there is a need to go beyond xAI. To reach a level of "explainable medicine" there is a crucial need for causability [39]. In the same way that usability encompasses measurements for the quality of use, causability encompasses measurements for the quality of explanations, e. g. the heatmaps produced by explainable AI methods. Causability can be seen as a property of "human intelligence", whereas explainability can be seen as the property of a "artificial intelligence".

The key to effective human-AI interaction and consequently the success of future human-AI interfaces lies in an efficient and consistent **mapping of explainability with causability** [40].

This "mapping" is about establishing connections and relationships between existing areas, so *not* about drawing a new map, but rather to identify similar areas in two completely different "maps". Effective and efficient mapping is necessary, but obviously not sufficient for understanding an explanation: Whether an explanation has been understood depends on other factors, including prior knowledge and expectations on the human side. Obviously, the effectiveness of an "explanation interface" depends on whether (and to what extent) the result of an explanation produced by an explainable AI method was understood by the human expert.

As we can imagine this is not trivial, because future Human-AI interfaces should allow a constant feedback, whether and to what extent something has been understood, or not. In a human-to-human interaction, this feedback is very much provided by facial expressions. Consequently, concepts of "emotion" [41], [42], "emotion detection" [43] and "emotional interfaces" [44] will become an important part of future conversational interfaces for explainable AI [45] and dialog systems [46]. Such features will become important for these future "explanation interfaces" or however we will call them. One very important aspect is to include a key component that has been used as a standard communication tool between doctors for centuries: *language*, i. e. to produce descriptive sentences based on *domain ontologies* to clarify the decision of deep learning classifiers, hence to augment the results with short quantified sentences of natural language [47].

Summarizing, humans will continue to play a special role in the AI pipeline in the foreseeable future, complementing capabilities of AI due to their genuine human abilities. The backbone of this approach is interactive machine learning [48], [49], [50] which adds the component of human expertise to AI processes by enabling them to re-enact and retrace the results on demand, e. g. let them check it for plausibility.

3 Towards Future Human-AI Interfaces for Multi-Modal Causability

We can recapitulate that in the medical domain we need to include a human-in-the-loop for several reasons: to complement AI with human expertise and conceptual knowledge, to augment the human with AI, and also to keep the human in control for social, ethical and legal reasons [51]. For all these reasons there is a pressing need for the design, development, and evaluation of new effective and efficient Human-AI interfaces. This challenges the Human-computer interaction (HCI) community: Design guidelines for human-AI interaction are already under way [52]. Moreover, general principles and design guidelines for interactive techniques have been discussed in the HCI community for decades [53], which are now becoming important again. Lastly, the quest for effective Human-AI interfaces was boosted recently by the xAI program of DARPA, where they explicitly emphasized the importance of *interactive "explanation interfaces"* [54] and where they emphasized that understanding (sensemaking) must be facilitated by *interactive* guided explanations. This is motivated by the fact that for a biomedical expert using AI, it is very important to be able to investigate the **independent** underlying factors which influenced the machine aided decision-making process, taking into account that we cannot always disentangle dependent factors. That said, decision paths defined by biomedical experts will capture *only a subset of the features available* to train machine learning models in medical AI. From this reduced feature set (multi-*omics and clinical parameters), it can be beneficial to build reference classification models based on decision trees which may reflect the biomedical decision process. Such decision trees can then act as a reference model, but most importantly, as a benchmark for the *reliability* of "black-box" AI models. We need to carefully study the accuracy of such reference models and to investigate their generalizability regarding heterogeneous patient profiles. In this context, disease subtypes can be derived. For this purpose, the development of new and the application of existing multi-view clustering algorithms [55] can be very helpful.

A very useful approach to combine various data in order to create comprehensive views of diseases or biological processes is *Similarity Network Fusion* (SNF) developed by Wang et al. [56]. This method solves the integration problem by constructing networks of samples for each available data type and fusing these into one single network that represents the underlying data. The increasing complexity of the biomedical domain and the introduction of new technologies enable investigations in arbitrarily high dimensional spaces, practically having millions of different properties (including genomics and proteomics, but also images, patient history, etc.). No single data type can, however, capture the complexity of all these factors which are relevant to understand a phenomenon, i. e. a disease. This calls for integrative methods that combine data from multiple technologies and provide a comprehensive and relevant system view [57], [5], [58].

An ideal method must be able to "answer" a biological or medical question, i. e. to identify important features and predict outcomes, by harnessing heterogeneous data across several dimensions of biological variation. Very useful in this context is *Neighbourhood based Multi-*omics clustering* (NEMO) [59]. NEMO can be applied to partial datasets without performing data imputation and works in three phases: First, an inter-patient similarity matrix is built for each *omics data; then the matrices of different *omics data are integrated into one single matrix; finally this network is clustered. A very recent approach is *Pathway Graph Kernel based Multi-Omics Approach for Patient Clustering* (*PAMOGK*) [60], that integrates multi-*omics patient data with existing biological knowledge on pathways. A graph kernel evaluates patient similarities based on a single molecular alteration type in the context of such a pathway, and to support multiple views, a multi-view kernel clustering is used. A measurement for the predictive power is the *Area under the Curve* (AUC). In the context of explainability/causability, however, only parts of the AUC are informative. This is mostly due to the fact that we are often confronted with imbalanced data in the biomedical domain [61]. Known alternatives such as the partial AUC cannot be fully interpreted, because they ignore some information about actual negatives. However, the recently developed concordant (partial-) pAUC is more useful [62] and may help to understand and interpret parts of the AUC.

Although the above mentioned models perform well, we are far from being able to use them within daily biomedical practice as long as *the underlying decision paths are not made visible*, and most importantly, *understandable and interpretable* for the end-user, because we still are confronted with the "black-box problem" [63]. Here we should note that the decision-making process can be seen as a sequence of steps in which the biomedical expert selects a path through a network of *plausible* events and actions. This goes back to the seminal work of Shortliffe et al. [64]: Nodes in this tree-shaped network are of two kinds: "decision nodes", where the expert can select from a set of actions, and "chance nodes", where the outcome cannot be directly controlled by the expert, but is a probabilistic response of the patient to some action taken. For example, a physician may choose to perform a certain test (decision node) but the occurrence or non-occurrence of complications may be largely a matter of statistical likelihood (chance node). By analyzing a difficult decision process before taking any action, it may be possible to delineate in advance all pertinent decision nodes and chance nodes along with all plausible outcomes, plus the paths by which these outcomes might be reached. To address this shortcoming, one possibility is to relate the multi-modal models, which are built on stochastic procedures only, to a biomedical expert's reference model. This requires investigating whether and to what extend the corresponding decision paths are reflected and/or covered. This can be done via *"what-if" (counterfactuals) requests* to the system, but also with additional *state-of-the-art* approaches that are widely used by the xAI community to date, for example, popular model-agnostic approaches such as DALEX [65], LIME [66], or, more recently, optiLIME [67]. All these approaches can be used for both global explainability (for model understanding) and for local explainability (for prediction understanding). Every explainer creates a numerical summary and a visual summary and allows for comparison of multiple models. To enhance the understandability for the domain expert, this can be augmented via *short quantified sentences on natural language* [47]. A big advantage of the counterfactual generation is that it can be considered as a post-hoc procedure which can act independent from any classifier [68]. The resulting counterfactuals can be modelled as a graph, where features are defined as nodes and the edges as combination of such features, which we call *"counterfactual paths"*. Initially, such a counterfactual graph may be generated in a purely data-driven manner. The distance between the counterfactuals (weighted edges) can be defined as in [69]. In an ideal setting, the automatic generation of the counterfactual paths is fully reflected by the leaf nodes of the medical decision trees [70]. To facilitate the human *interaction* with the multi-modal machine learning model opens new ways of interactive Human-AI interface, supporting both explainability and causability. Here, the guiding idea is that the biomedical experts are empowered to ask questions ("why are the cells smaller and closer together") and also counterfactual "what-if" questions ("what if the cells are slightly bigger"). Here a chance is to derive *simple-to-understand* decision trees derived from the graph, which itself can be derived by a decision forest classifier comprising multiple trees derived from the counterfactual classes. Recent work has shown how to efficiently reduce such a decision forest to a single decision tree [71], [72] from which counterfactuals can be easily observed based on the leaf nodes. Here. the human in the loop will have the opportunity to study this consensus decision tree. and the domain expert will be able to adopt the modifications to the counterfactual graph accordingly (feedback-loop). It is necessary to visualize relevant input features as well as the underlying explanatory factors, the "decision network". This is of course a non-trivial task because such a visualization has to be optimized to a) the human visual perception capability, b) the prior knowledge of the human, and c) the context of the workflow. This calls for *flexible* interfaces, taking into consideration existing methods, algorithms and tools [73], [74], from a) post-hoc interpretable models and systems, which aim to provide local explanations for a specific decision and making it reproducible on demand (instead of explaining the whole model behaviour), over b) ante-hoc models, which are interpretable by design which includes so called glass-box approaches

[50]. Of course, a lot of further research in the real-world is needed regarding the technical parameters' robustness and explainability.

4 Conclusion

Thanks to the great progress in statistical learning, we are experiencing an AI renaissance. Available and practical useable deep learning approaches achieve a performance that is beyond human level performance – even in the medical domain. This is a great success and there is no question that AI will become very important for medicine. Especially, when considering what humans are *not* able to do – but AI can. Nevertheless: correlation is not causation and contemporary AI models have become so complex that they are considered as "black-box" approaches. This makes it hard for domain experts to understand why a certain result has been achieved. The xAI community has developed a variety of methods for making such approaches transparent. This constitutes a promising first step, but while xAI deals with the implementation of transparency and traceability in statistical black-box machine learning methods in the medical domain, there is a pressing need to go beyond xAI: to reach a level of explainable medicine we need causability, which encompasses measurements for the quality of explanations produced by xAI methods (e. g. heatmaps). Here, very important is the human in the loop, because (sometimes) a human expert is necessary to add contextual understanding and experience. This, in turn, requires new ***interactive*** human-AI interfaces, especially in the medical domain, in which many different modalities contribute to a result. To support future "explainable medicine", we therefore need multi-modal causability. That said, we need interactive Human-AI interfaces which enable a domain expert to ask questions to understand why a machine came up with a result, and to ask "what-if" questions (counterfactuals) to gain insight into the underlying independent explanatory factors of a result. Overall, intensive research and development on an international level will thus be necessary in to make AI even more successful and to use medical AI effectively for the benefit of human health.

Acknowledgment: I am very grateful for the invitation by the workshop chairs to provide a keynote for the User-Centered Artificial Intelligence (UCAI) workshop [75] in the context of the Mensch & Computer 2020 conference. This paper is based on the given lecture. I am grateful for the reviewer's comments on the manuscript which helped to improve this paper.

Funding: Parts of this work have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826078 "Feature Cloud" and parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 "a reference model for explainable Artificial Intelligence".

Glossary

Bias inability of an algorithm to represent the true relationship; High bias can cause an algorithm to miss the relevant relations between features and output.

Causal inference the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect. In medicine we call the science of why things occur etiology (the study of the causation of pathologies).

Causability is a property of a human (natural intelligence) and a measurement for the degree of human understanding. Future human-centered AI interfaces must ensure a mapping between explainability and causability, i. e. between explanations generated by an xAI method and the prior knowledge of the human.

Counterfactual a hypothesis that is contrary to the facts (similar to counterexample), or a hypothetical state of the world, used to assess the impact of an action in the real-world, or a conditional statement in which the conditional clause is false, as "what-if" – this is very important to enable a human expert to ask such questions in human-centered AI interfaces.

Counterexample an exception of a proposed general rule or law and appears as an example which disproves a general statement made.

Explainability motivated by the opaqueness of so called "black-box" approaches, the ability to provide an explanation on why a machine decision has been reached, technically by highlighting the factors which contributed to the classification result.

Explanation set of statements to describe a given set of facts to clarify causality, context and consequences thereof; it is a core topic of knowledge discovery involving "why" questions, and "what-if" questions (counterfactuals).

Explicit Knowledge can be explained, e. g. by articulating it via natural language etc. and can be shared with other people.

European General Data Protection Regulation (EU GDPR) Regulation EU 2016/679 – see the EUR-Lex 32016R0679, will make black-box approaches difficult to use, because they often are not able to explain why a decision has been made.

Ground truth generally information provided by direct observation (i. e. empirical evidence) instead of provided by inference. For us it is the gold standard, i. e. the ideal expected result (100 % true).

KANDINSKY-Patterns an exploration environment used as "a swiss knife for the study of explainability " [76]– see https://www.youtube.com/watch?v=UuiV0icAlRs

Robustness a characteristic of a biological system (also called biological or genetic robustness is the persistence of a certain characteristic or trait in a system under perturbations or conditions of uncertainty.

Tacit Knowledge Knowledge gained from personal experience that is even more difficult to express than implicit knowledge.

References

[1] Minsky, Marvin 1961. Steps Towards Artificial Intelligence. *Proceedings of the Institute of Radio Engineers*, 49, (1), 8–30, doi:10.1109/jrproc.1961.287775.

[2] Vapnik, Vladimir N. 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10, (5), 988–999, doi:10.1109/72.788640.

[3] Hastie, Trevor, Tibshirani, Robert & Friedman, Jerome 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition*, New York, Springer, doi:10.1007/978-0-387-84858-7.

[4] Lecun, Yann, Bengio, Yoshua & Hinton, Geoffrey 2015. Deep learning. *Nature*, 521, (7553), 436–444, doi:10.1038/nature14539.

[5] Holzinger, Andreas, Haibe-Kains, Benjamin & Jurisica, Igor 2019. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*, 46, (13), 2722–2730, doi:10.1007/s00259-019-04382-9.

[6] Regitnig, Peter, Mueller, Heimo & Holzinger, Andreas 2020. Expectations of Artificial Intelligence in Pathology. *Springer Lecture Notes in Artificial Intelligence LNAI 12090*. Cham: Springer, pp. 1–15, doi:10.1007/978-3-030-50402-1-1.

[7] Hosny, Ahmed, Parmar, Chintan, Quackenbush, John, Schwartz, Lawrence H. & Aerts, Hugo J.W.L. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18, (8), 500–510, doi:10.1038/s41568-018-0016-5.

[8] Holzinger, Andreas, Stocker, Christof, Ofner, Bernhard, Prohaska, Gottfried, Brabenetz, Alberto & Hofmann-Wellenhof, Rainer 2013. Combining HCI, Natural Language Processing, and Knowledge Discovery – Potential of IBM Content Analytics as an assistive technology in the biomedical domain. *Springer Lecture Notes in Computer Science LNCS, vol. 7947*, Heidelberg, Berlin, New York. Springer, 13–24, doi:10.1007/978-3-642-39146-0_2.

[9] Rahim, Sarni Suhaila, Palade, Vasile, Almakky, Ibrahim & Holzinger, Andreas 2019. Detection of Diabetic Retinopathy and Maculopathy in Eye Fundus Images Using Deep Learning and Image Augmentation. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, pp. 114–127, doi:10.1007/978-3-030-29726-8_8.

[10] Jean-Quartier, Claire, Jeanquartier, Fleur, Jurisica, Igor & Holzinger, Andreas 2018. In silico cancer research towards 3R. *Springer/Nature BMC cancer*, 18, (1), 408, doi:10.1186/s12885-018-4302-0.

[11] Yuan, Hao, Tang, Jiliang, Hu, Xia & Ji, Shuiwang 2020. XGNN: Towards Model-Level Explanations of Graph Neural Networks. In: Liu, Yan & Gupta, Rajesh (eds.) *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*. San Diego (CA): ACM. 430–438, doi:10.1145/3394486.3403085.

[12] Lakkaraju, Himabindu, Kamar, Ece, Caruana, Rich & Leskovec, Jure 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.

[13] Lakkaraju, Himabindu, Kamar, Ece, Caruana, Rich & Leskovec, Jure. Faithful and customizable explanations of black box models. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), 2019. 131–138.

[14] Rudin, Cynthia 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, (5), 206–215, doi:10.1038/s42256-019-0048-x.

[15] Ras, Gabrielle, Haselager, Pim & Van Gerven, Marcel 2018. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. arXiv:1803.07517.

[16] Ribeiro, Marco Tulio, Singh, Sameer & Guestrin, Carlos 2016. Model-Agnostic Interpretability of Machine Learning. arXiv:1606.05386.

[17] Faust, Kevin, Bala, Sudarshan, Van Ommeren, Randy, Portante, Alessia, Al Qawahmed, Raniah, Djuric, Ugljesa & Diamandis, Phedias 2019. Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. *Nature Machine Intelligence*, 1, (7), 316–321, doi:10.1038/s42256-019-0068-6.

[18] Mccarthy, John 2007. From here to human-level AI. *Artificial Intelligence*, 171, (18), 1174–1182, doi:10.1016/j.artint.2007.10.009.

[19] Lake, Brenden M., Ullman, Tomer D., Tenenbaum, Joshua B. & Gershman, Samuel J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, (e253), doi:10.1017/S0140525X16001837.

[20] Chen, Xi, Duan, Yan, Houthooft, Rein, Schulman, John, Sutskever, Ilya & Abbeel, Pieter 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Lee, Daniel, Sugiyama, Mashahi, Luxburg, Ulrike V., Guyon, Isabelle & Garnett, Roman (eds.), *Advances in neural information processing systems*. Barcelona: NIPS foundation. 2172–2180.

[21] Finlayson, Samuel G., Bowers, John D., Ito, Joichi, Zittrain, Jonathan L., Beam, Andrew L. & Kohane, Isaac S. 2019. Adversarial attacks on medical machine learning. *Science*, 363, (6433), 1287–1289, doi:10.1126/science.aaw4399.

[22] Narayanan, Menaka, Chen, Emily, He, Jeffrey, Kim, Been, Gershman, Sam & Doshi-Velez, Finale 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. arXiv:1802.00682.

[23] Goebel, Randy, Chander, Ajay, Holzinger, Katharina, Lecue, Freddy, Akata, Zeynep, Stumpf, Simone, Kieseberg, Peter & Holzinger, Andreas 2018. Explainable AI: the new 42? *Springer Lecture Notes in Computer Science LNCS 11015*. Cham: Springer, pp. 295–303, doi:10.1007/978-3-319-99740-7-21.

[24] Holzinger, Andreas, Kieseberg, Peter, Weippl, Edgar & Tjoa, A Min 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. *Springer Lecture Notes in Computer Science LNCS 11015*. Cham: Springer, pp. 1–8, doi:10.1007/978-3-319-99740-7-1.

[25] Samek, Wojciech, Montavon, Gregorie, Vedaldi, Andrea, Hansen, Lars Kai & Müller, Klaus-Robert, (eds.) 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, Cham: Springer Nature, doi:10.1007/978-3-030-28954-6.

[26] Arrieta, Alejandro Barredo, Díaz-Rodríguez, Natalia, Del Ser, Javier, Bennetot, Adrien, Tabik, Siham, Barbado, Alberto, García, Salvador, Gil-López, Sergio, Molina, Daniel, Benjamins, Richard, Chatila, Raja & Herrera, Francisco 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115, doi:10.1016/j.inffus.2019.12.012.

[27] Holzinger, Andreas, Biemann, Chris, Pattichis, Constantinos S. & Kell, Douglas B. 2017. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923.

[28] Bach, Sebastian, Binder, Alexander, Müller, Klaus-Robert & Samek, Wojciech 2016. Controlling explanatory heatmap resolution and semantics via decomposition depth. *2016 IEEE International Conference on Image Processing (ICIP)*. Phoenix (AZ): IEEE. 2271–2275, doi:10.1109/ICIP.2016.7532763.

[29] Sturm, Werner, Schaefer, Till, Schreck, Tobias, Holzinger, Andeas & Ullrich, Torsten 2015. Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps. In: Borgo, Rita & Turkay, Cagatay (eds.) *EG UK Computer Graphics & Visual Computing CGVC 2015*. University College London (UCL): Euro Graphics (EG). 77–84, doi:10.2312/cgvc.20151247.

[30] Montavon, Grégoire 2019. Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison. In: Samek, Wojciech, Montavon, Grégoire, Vedaldi, Andrea, Hansen, Lars Kai & Müller, Klaus-Robert (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer International Publishing, pp. 253–265, doi:10.1007/978-3-030-28954-6-13.

[31] Lapuschkin, Sebastian, Binder, Alexander, Montavon, Gregoire, Müller, Klaus-Robert & Samek, Wojciech 2016. The LRP toolbox for artificial neural networks. *The Journal of Machine Learning Research (JMLR)*, 17, (1), 3938–3942.

[32] Montavon, Gregoire, Lapuschkin, Sebastian, Binder, Alexander, Samek, Wojciech & Müller, Klaus-Robert 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65, 211–222, doi:10.1016/j.patcog.2016.11.008.

[33] Montavon, Gregoire, Samek, Wojciech & Müller, Klaus-Robert 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, (2), 1–15, doi:10.1016/j.dsp.2017.10.011.

[34] Schnake, Thomas, Eberle, Oliver, Lederer, Jonas, Nakajima, Shinichi, Schütt, Kristof T., Müller, Klaus-Robert & Montavon, Grégoire 2020. XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. arXiv:2006.03589.

[35] Zeiler, Matthew D., Krishnan, Dilip, Taylor, Graham W. & Fergus, Rob 2010. Deconvolutional networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), IEEE, 2528–2535, doi:10.1109/CVPR.2010.5539957.

[36] Zeiler, Matthew D., Taylor, Graham W. & Fergus, Rob. 2011. Adaptive deconvolutional networks for mid and high level feature learning. IEEE International Conference on Computer Vision (ICCV), IEEE, 2018–2025, doi:10.1109/ICCV.2011.6126474.

[37] Zeiler, Matthew D. & Fergus, Rob 2014. Visualizing and understanding convolutional networks. In: Fleet, David, Pajdla, Tomas, Schiele, Bernt & Tuytelaars, Tinne, (eds.), *ECCV, Lecture Notes in Computer Science LNCS 8689*. Cham: Springer, pp. 818–833, doi:10.1007/978-3-319-10590-1-53.

[38] Kim, Been, Wattenberg, Martin, Gilmer, Justin, Cai, Carrie, Wexler, James & Viegas, Fernanda. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673–2682.

[39] Holzinger, Andreas, Langs, Georg, Denk, Helmut, Zatloukal, Kurt & Müller, Heimo 2019. Causability and Explainability of Artificial Intelligence in Medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9, (4), 1–13, doi:10.1002/widm.1312.

[40] Holzinger, Andreas, Carrington, Andre & Müller, Heimo 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. *KI – Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt*, 34, (2), 193–198, doi:10.1007/s13218-020-00636-z.

[41] Mayer, John D. & Geher, Glenn 1996. Emotional intelligence and the identification of emotion. *Intelligence*, 22, (2), 89–113, doi:10.1016/S0160-2896(96)90011-2.

[42] Picard, R. W., Vyzas, E. & Healey, J. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, (10), 1175–1191.

[43] Stickel, Christian, Ebner, Martin, Steinbach-Nordmann, Silke, Searle, Gig & Holzinger, Andreas 2009. Emotion Detection: Application of the Valence Arousal Space for Rapid Biological Usability Testing to Enhance Universal Access. In: Stephanidis, Constantine (ed.), *Universal Access in Human-Computer Interaction. Addressing Diversity, Lecture Notes in Computer Science, LNCS vol. 5614*. Berlin, Heidelberg: Springer, pp. 615–624, doi:10.1007/978-3-642-02707-9-70.

[44] Picard, Rosalind W., Wexelblat, Alan & Nass, Clifford I. 2002. Future interfaces: social and emotional. *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. 698–699.

[45] Jentzsch, Sophie F., Höhn, Sviatlana & Hochgeschwender, Nico 2019. Conversational Interfaces for Explainable AI: A Human-Centred Approach. *International Workshop on*

Explainable, Transparent Autonomous Agents and Multi-Agent Systems. Springer. 77–92.

[46] Merdivan, Erinc, Singh, Deepika, Hanke, Sten & Holzinger, Andreas 2019. Dialogue Systems for Intelligent Human Computer Interactions. *Electronic Notes in Theoretical Computer Science*, 343, 57–71, doi:10.1016/j.entcs.2019.04.010.

[47] Hudec, Miroslav, Bednárová, Erika & Holzinger, Andreas 2018. Augmenting Statistical Data Dissemination by Short Quantified Sentences of Natural Language. *Journal of Official Statistics (JOS)*, 34, (4), 981, doi:10.2478/jos-2018-0048.

[48] Holzinger, Andreas 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? *Brain Informatics*, 3, (2), 119–131, doi:10.1007/s40708-016-0042-6.

[49] Holzinger, Andreas, Plass, Markus, Holzinger, Katharina, Crisan, Gloria Cerasela, CPintea, Camelia-M. & CPalade, Vasile 2016. Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach. *Springer Lecture Notes in Computer Science LNCS 9817*. Heidelberg, Berlin, New York: Springer, pp. 81–95, doi:10.1007/978-3-319-45507-56.

[50] Holzinger, Andreas, Plass, Markus, Kickmeier-Rust, Michael, Holzinger, Katharina, Crişan, Gloria Cerasela, Pintea, Camelia-M. & Palade, Vasile 2019. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, 49, (7), 2401–2414, doi:10.1007/s10489-018-1361-5.

[51] Schneeberger, David, Stoeger, Karl & Holzinger, Andreas 2020. The European legal framework for medical AI. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Fourth IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Proceedings*. Cham: Springer, pp. 209–226, doi:10.1007/978-3-030-57321-8-12.

[52] Amershi, Saleema, Weld, Dan, Vorvoreanu, Mihaela, Fourney, Adam, Nushi, Besmira, Collisson, Penny, Suh, Jina, Iqbal, Shamsi, Bennett, Paul N. & Inkpen, Kori. Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019 Glasgow. ACM, doi:10.1145/3290605.3300233.

[53] Ziegler, Jürgen 1996. Interactive techniques. *ACM Computing Surveys (CSUR)*, 28, (1), 185–187, doi:10.1145/234313.234392.

[54] Gunning, David & Aha, David W. 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, 40, (2), 44–58.

[55] Rappoport, Nimrod & Shamir, Ron 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46, (20), 10546–10562, doi:10.1093/nar/gky889.

[56] Wang, Bo, Mezlini, Aziz M., Demir, Feyyaz, Fiume, Marc, Tu, Zhuowen, Brudno, Michael, Haibe-Kains, Benjamin & Goldenberg, Anna 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11, (3), 333–340, doi:10.1038/nMeth.2810.

[57] Holzinger, Andreas & Jurisica, Igor 2014. Knowledge Discovery and Data Mining in Biomedical Informatics: The future is in Integrative, Interactive Machine Learning Solutions. In: Holzinger, Andreas & Jurisica, Igor (eds.), *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401*. Heidelberg, Berlin: Springer, pp. 1–18, doi:10.1007/978-3-662-43968-5_1.

[58] Zitnik, Marinka, Nguyen, Francis, Wang, Bo, Leskovec, Jure, Goldenberg, Anna & Hoffman, Michael M. 2019. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50, (10), 71–91, doi:10.1016/j.inffus.2018.09.012.

[59] Rappoport, Nimrod & Shamir, Ron 2019. NEMO: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35, (18), 3348–3356, doi:10.1093/bioinformatics/btz058.

[60] Tepeli, Yasin Ilkagan, Ünal, Ali Burak, Akdemir, Furkan Mustafa & Tastan, Oznur 2020. PAMOGK: A Pathway Graph Kernel based Multi-Omics Approach for Patient Clustering. *Bioinformatics*, btaa655, doi:10.1093/bioinformatics/btaa655.

[61] Lopez, V., Fernandez, A., Garcia, S., Palade, V. & Herrera, F. 2013. An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information Sciences*, 250, 113–141, doi:10.1016/j.ins.2013.07.007.

[62] Carrington, Andre M., Fieguth, Paul W., Qazi, Hammad, Holzinger, Andreas, Chen, Helen H., Mayr, Franz & Manuel, Douglas G. 2020. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *Springer/Nature BMC Medical Informatics and Decision Making*, 20, (1), 1–12, doi:10.1186/s12911-019-1014-6.

[63] Bhatt, Umang, Xiang, Alice, Sharma, Shubham, Weller, Adrian, Taly, Ankur, Jia, Yunhan, Ghosh, Joydeep, Puri, Ruchir, Moura, José Mf & Eckersley, Peter. Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020. 648–657, doi:10.1145/3351095.3375624.

[64] Shortliffe, Edward H., Buchanan, Bruce G. & Feigenbaum, Edward A. 1979. Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. *Proceedings of the IEEE*, 67, (9), 1207–1224, doi:10.1109/PROC.1979.11436.

[65] Biecek, Przemysław 2018. DALEX: explainers for complex predictive models in R. *The Journal of Machine Learning Research*, 19, (1), 3245–3249.

[66] Ribeiro, Marco Tulio, Singh, Sameer & Guestrin, Carlos 2016. Why should i trust you?: Explaining the predictions of any classifier. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*. San Francisco (CA): ACM. 1135–1144, doi:10.1145/2939672.2939778.

[67] Visani, Giorgio, Bagli, Enrico & Chesani, Federico 2020. OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms. arXiv:2006.05714.

[68] Mothilal, Ramaravind K., Sharma, Amit & Tan, Chenhao 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In: Hildebrandt, Mireille, Castillo, Carlos, Celis, Elisa, Ruggieri, Salvatore, Taylor, Linnet & Zanfir-Fortuna, Gabriela (eds.) *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT 2020)*. Barcelona: ACM. 607–617, doi:10.1145/3351095.3372850.

[69] Mahajan, Divyat, Tan, Chenhao & Sharma, Amit 2019. Preserving causal constraints in counterfactual explanations

for machine learning classifiers. arXiv:1912.03277.

[70] Karimi, Amir-Hossein, Von Kügelgen, Julius, Schölkopf, Bernhard & Valera, Isabel 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. arXiv:2006.06831.

[71] Fernández, Rubén R., De Diego, Isaac Martín, Aceña, Víctor, Fernández-Isabel, Alberto & Moguerza, Javier M. 2020. Random Forest explainability using counterfactual sets. *Information Fusion*, 63, (11), 196–207, doi:10.1016/j.inffus.2020.07.001.

[72] Sagi, Omer & Rokach, Lior 2020. Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*, 61, 124–138, doi:10.1016/j.inffus.2020.03.013.

[73] Cvek, Urska, Trutschl, Marian & Clifford, John 2010. Neural-network enhanced visualization of high-dimensional data. *Self-Organizing Maps*. InTech, pp. 145–165.

[74] Trutschl, Marjan, Kilgore, Phillip C. & Cvek, Urska. Self-Organization in Parallel Coordinates. International Conference on Artificial Neural Networks, 2013. Springer, 351–358.

[75] Augstein, Mirjam, Buschek, Daniel, Herder, Eelco, Loepp, Benedikt, Yigitbas, Enes & Ziegler, Jürgen (eds.) 2020. *UCAI 2020: Workshop on User-Centered Artificial Intelligence*, doi:10.18420/muc2020-ws111.

[76] Holzinger, Andreas, Kieseberg, Peter & Müller, Heimo 2020. KANDINSKY Patterns: A Swiss-Knife for the Study of Explainable AI. *ERCIM News*, (120), 41–42.

Bionotes

Andreas Holzinger
Human-Centered AI Lab, Institute for Medical Informatics & Statistics, Medical University Graz, Graz, Austria
xAI Lab, Alberta Machine Intelligence Institute, Edmonton, Canada
andreas.holzinger@medunigraz.at

Andreas Holzinger promotes a synergistic approach of Human-Centered AI (HCAI) to put the human-in-control of AI to align it with human values, privacy, security and safety, motivated by efforts to improve human health. He is lead of the Human-Centered AI Lab (Holzinger Group) at the Medical University Graz and Visiting Professor for explainable AI at the University of Alberta, Edmonton, Canada. Since 2016 Andreas is teaching machine learning in health informatics at Vienna University of Technology. He serves as consultant for the Canadian, US, UK, Swiss, French, Italian and Dutch governments, for the German Excellence Initiative, and as national expert in the European Commission. He is in the advisory board of the Artificial Intelligence Strategy “AI Made in Germany 2030” of the German Federal Government. human-in-the-loop. Andreas obtained a Ph.D. in Cognitive Science from Graz University in 1998 and his second Ph.D. in Computer Science from TU Graz in 2003. He serves as Austrian Representative for Artificial Intelligence in IFIP TC 12 and in this position is organizer of the IFIP Cross-Domain Conference “Machine Learning & Knowledge Extraction (CD-MAKE)”, and is member of IFIP WG 12.9 Computational Intelligence. Andreas Holzinger was elected a full member of the Academia Europea – the European Academy of Sciences, in 2019 in the section informatics. More information: https://www.aholzinger.at