European Research and Innovation

# KANDINSKY Patterns: A Swiss-Knife for the Study of Explainable AI

by Andreas Holzinger (Medical University Graz and Alberta Machine Intelligence Institute Edmonton, Canada), Peter Kieseberg (University of Applied Sciences, St.Poelten) and Heimo Müller (Medical University Graz)

***Kandinsky Patterns are mathematically describable, simple, self-contained, hence controllable test datasets for the development, validation and training of explainability in artificial intelligence (AI) and machine learning (ML).***

In machine learning we design, we develop, we test, and we evaluate algorithms, which can learn from data and extract knowledge and make predictions. This is very helpful for decision support systems and decision making, particularly in the medical domain. One family of machine learning algorithms is deep learning, which is now extremely successful. Deep learning reaches human-level performance in classification tasks even in the medical domain, with a classification performance of 92% and better. However, the big problem here is its "black-box" nature, i.e., a medical expert is currently unable to retrace, replicate and understand the underlying explanatory factors of why the 92% have been achieved, for instance. Particularly in medicine, the question of why is often more important than the classification result itself [1], therefore the field of "explainable AI" is becoming more and more important [2].

Human experts can explain certain results very well, because humans are able to use abstract concepts. Causality and concept learning are important to understand how humans extract so much information, often from just a few data points, and to contrast this with machine learning which is now addressing "explainable AI". For the study of explainable AI, we have designed and developed an experimental explanation environment for testing explainability concepts of both human intelligence and artificial intelligence. We made this environment fully accessible for the international machine learning community and called it Kandinsky Patterns, in memory of the great painter Wassily Kandinsky. But there is another story behind these patterns: In the past we have observed how pathologists work [1]. Pathology is a very interesting medical specialty; pathologists explain and interpret geometric objects and geometric structures (Figure 1). They even speak of geometrical architectures, and they observe and interpret shapes, objects, colour, similarity, Gestalt phenomena, etc., and finally come up with an explanation in written form - the diagnosis. For a number of reasons, this process is very difficult for machine learning algorithms.

The Kandinsky Patterns enable the study of such phenomena and to test, benchmark and evaluate machine learning algorithms under mathematically strictly controllable conditions [L1], [L2], but which at the same time are accessible and

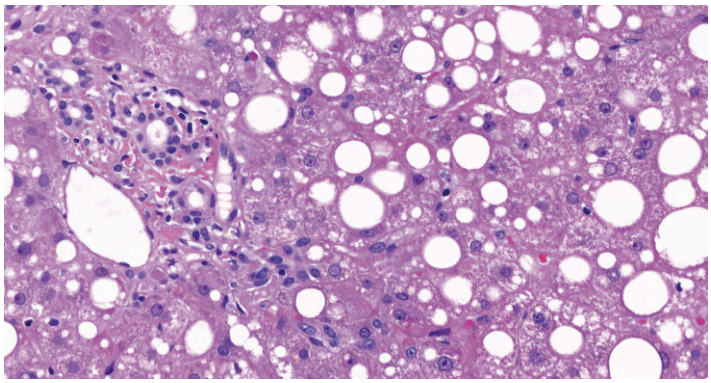

*Figure 1: A typical interpretation task in pathology: The (human) pathologist interprets and explains geometrical architectures, shapes, objects, colour, similarity, etc., and produces an explanation – the medical report [1].*

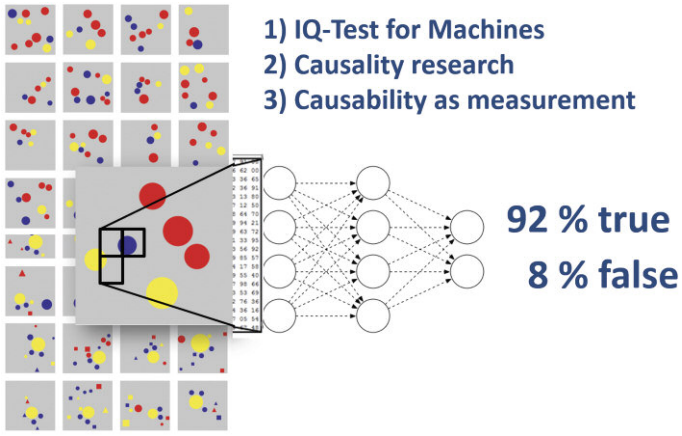


*Figure 2: Kandinksy Patterns as an experimental environment for explainable AI – to answer the question of why the classificatory came up with 92% [L2].*

understandable for human observers. It is also possible to produce a ground truth. That means, we can produce images that the tested classifier should evaluate as true, thereby allowing various ways of testing machine learning algorithms. Thus, Kandinsky Patterns can be used as a kind of intelligence test for algorithms [3]. Kandinsky Patterns allow validation of results of arbitrary machine learning algorithms, but most of all, explainable AI methods can be tested, evaluated and further developed based on these experiments. This even allows for the development of completely new explanation methods. Moreover, with our Kandinsky Generator we can produce " false patterns" that can be used to test the robustness of algorithms in a controlled setting. This will be extremely important in the future, as adversarial examples have already demonstrated their potential in attacking security mechanisms applied in various domains, especially medical environments. Last, but not least, Kandinsky Patterns can be used to produce "counterfactuals" – the "what if", which is difficult to handle for both humans and machines - but can provide new insights into the behaviour of explanation methods.

In conclusion, Kandinsky Patterns (Figure 2) can be used as "IQ-Test for machines", for causality research and to evaluate explainable AI methods, i.e. to use causality as measurement [1].

**Links:**

[L1] https://human-centered.ai/project/kandinsky-patterns/
[L2] https://www.youtube.com/watch?v=UuiV0icAlRs

**References:**

[1] A. Holzinger, et al.: 2019. Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, (4), 2019, doi:10.1002/widm.1312
[2] R. Goebel, et al.:"Explainable AI: the new 42?, in International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 295-303), 2018. doi: doi:10.1007/978-3-319-99740-7_21
[3] A. Holzinger, M. Kickmeier-Rust, H. Müller: "KANDINSKY Patterns as IQ-Test for Machine Learning", in International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 1-14). Doi:10.1007/978-3-030-29726-8_1

**Please contact:**

Andreas Holzinger
exAI Lab, Alberta Machine Intelligence Institute, Edmonton, Canada and Medical University Graz, Austria
andreas.holzinger@medunigraz.at