Reg. No: 2021/43/O/ST6/00347; Principal Investigator: dr hab. inż. Przemysław Biecek

ARES: Attack-Resistant Explanations toward Secure and trustworthy AI

Machine learning explainability, fairness, robustness, and security are key elements of trustworthy Artificial Intelligence, an area of strategic importance. In this context, the main goals of the ARES project are:

1. Develop **adversarial attacks** on state-of-the-art explanations to investigate vulnerabilities and limitations of the existing explainability and fairness approaches in machine learning.
2. Introduce novel **robust explanations** that are stable against manipulation and intuitive to evaluate.

We target to progress explainable machine learning toward a secure and trustworthy adoption of AI solutions.

Questions and hypotheses: Specifically, we aim to address the following research questions and hypotheses: What are the critical limitations of state-of-the-art explanations? How to detect the potential manipulation of explanations? State-of-the-art explanations for machine learning models trained on tabular data are not robust nor trustworthy in the context of adversarial attacks on explanations. What to improve toward developing robust explanations? How to evaluate explanations in the context of an adversary? Robust explanations are of novel quality and resist against the potential attacks.

Impact: Despite rapid development in recent years, the research field of explainable machine learning is held back by a lack of consideration for the security and evaluation of novel methods and tools. Historically, it has been so with scenarios considering adversary and security in machine learning. Achieving the first goal primarily impacts various domains of research, which currently use (and explain) black-box models for knowledge discovery and decision-making, by highlighting vulnerabilities and limitations of their explanations. Achieving the second goal impacts more the broad machine learning domain as it aims at improving state-of-the-art by introducing robust explanations toward secure and trustworthy AI.

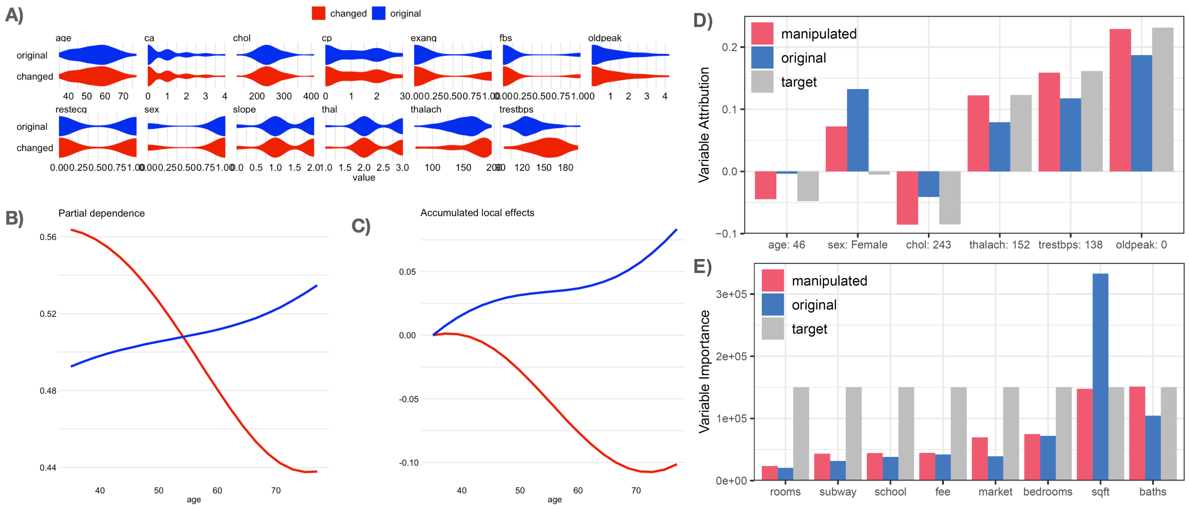

Figure 1: In preliminary research [1–3], we effectively manipulate (change) explanations for tabular data in real-world scenarios by exploiting their vulnerability to **A)** data poisoning, e.g. **B)** partial dependence, **C)** accumulated local effects, **D)** SHAP attribution, and **E)** SHAP importance, which is for a financial use-case. The blue lines and bars denote original explanations, while the manipulated ones are in red.

References

[1] H. Baniecki and P. Biecek. Manipulating SHAP via Adversarial Data Perturbations (Student Abstract). In *(to appear) AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[2] H. Baniecki, W. Kretowicz, and P. Biecek. Fooling Partial Dependence via Data Poisoning. *arXiv preprint arXiv:2105.12837*, 2021.

[3] K. Woźnica, K. Pękala, H. Baniecki, W. Kretowicz, E. Sienkiewicz, and P. Biecek. Do not explain without context: addressing the blind spot of model explanations. *arXiv preprint arXiv:2105.13787*, 2021.