Robustness of Explainable Artificial Intelligence in Industrial Process Modelling

Benedikt Kantz, Clemens Staudinger, Christoph Feilmayr, Johannes Wachlmayr, Alexander Haberl, Stefan Schuster, Franz Pernkopf

Signal Processing and Speech Communication Laboratory - Graz, voestalpine Stahl GmbH - Linz

At a glance

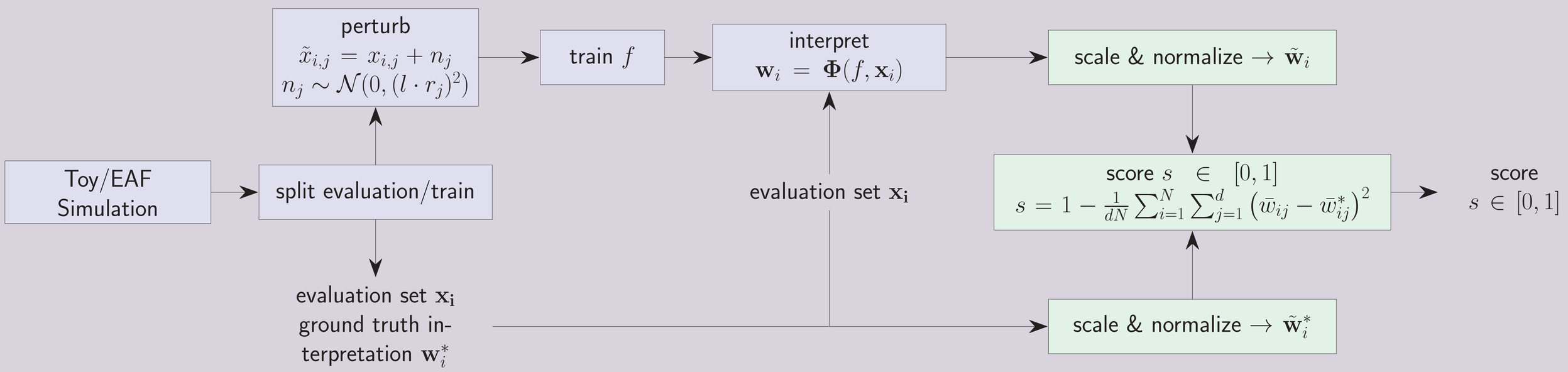
- Problem: Performance of eXplainable Artificial Intelligence (XAI) methods not evaluated in noisy settings
- Approach: Evaluation pipeline and comparison of explanations to ground truth effects
- Results: Explainer performance depends on Machine Learning (ML) model performance, robust XAI methods consider many gradients of a robust ML model.

Problem & Challenges

- XAI & *effect modeling* is key for industrial processes (*digital surrogates*) to understand the models and the perturbations of the inputs
- Noise robustness needs to be quantified
- Ground truth effect $\mathbf{w}_i^*$ not available in real-world data $\rightarrow$ **simulated datasets with ground truth!**
- Scoring for XAI methods difficult $\rightarrow$ **evaluation method proposed!**
- Different kinds of XAI methods
  - *Effect-based* methods: Gradient, SG, ALE-kNN
  - *Attribution-based* methods: LIME, SHAP

Our Evaluation methodology

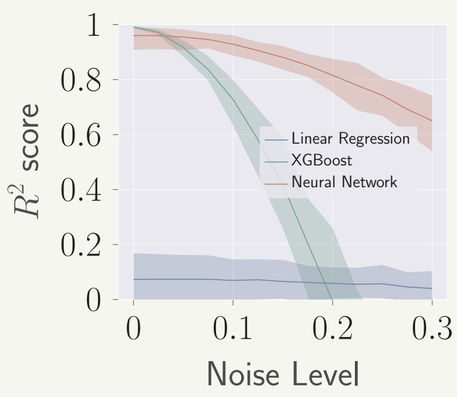
- Score scaling & Alignment
- Artififcially perturb dataset using noise $n_j \sim \mathcal{N}(0, (l \cdot r_j)^2)$ based on data range r_j
- Train model $f(\mathbf{x})$
- Infer local interpretations $\mathbf{w}_i = \mathbf{\Phi}(f, \mathbf{x}_i)$
- Calculate score $s \in [0, 1]$



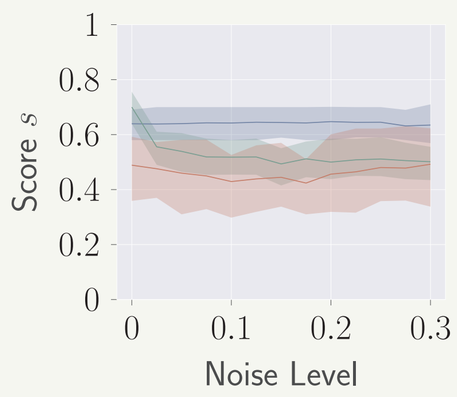
Results

- Toy dataset: Polynom $f(x_1, x_2) = k_1 x_1^2 + k_2 x_2^2 + k_3 x_1 x_2 + k_4 x_1 + k_5 x_2 + k_6$
  - Generate 1000 samples
  - Calculate ground truth $\mathbf{w}^*$ using automatic differentiation
  - R^2 scored & score s with varying levels of noise and different combinations of explainers and ML models.

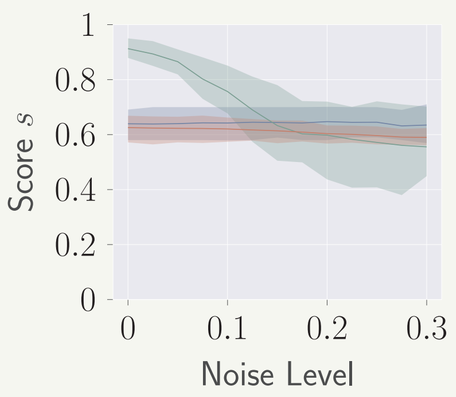
(a) R^2 Score



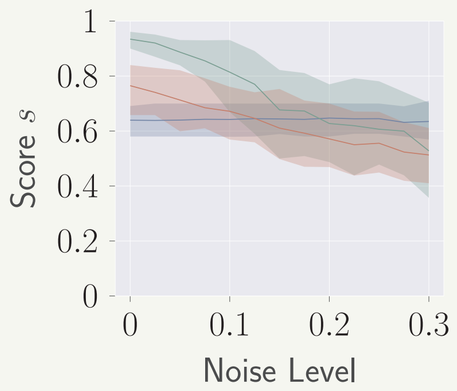
(b) Gradient



(c) Smooth Gradients (SG)



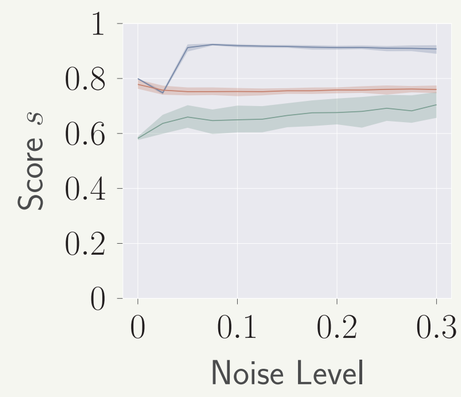
(d) Averaged Local Effects (ALE)-k-Nearest Neighbors (kNN)



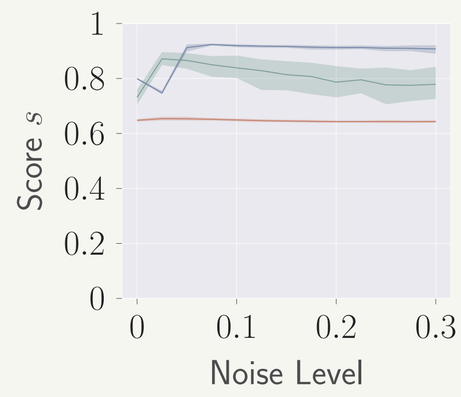
- Electric Arc Furnace (EAF) simulation
  - Relevancy: Sustainable alternative to blast furnaces, well-researched chemical & electrical problem
  - Chemical simulation for different input parameters; observed auxiliary parameters & target value (carbon in tapped steel)
  - Calculate ground truth $\mathbf{w}^*$ using automatic differentiation through simulation
  - R^2 scored & score s with varying levels of noise and different combinations of explainers and ML models.

(a) R^2 Score

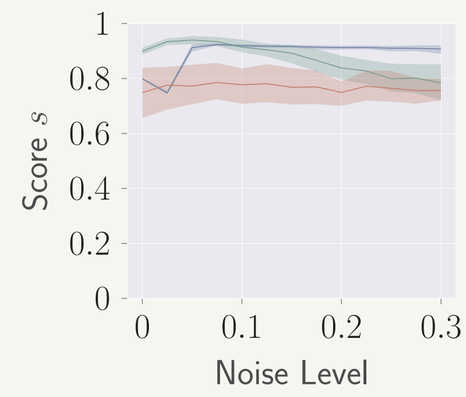
(b) Gradient



(c) SG



(d) ALE-kNN



https://www.spsc.tugraz.at/ {benedikt.kantz,pernkopf}@tugraz.at