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 University of Technology, voestalpine AG - Linz

At a glance

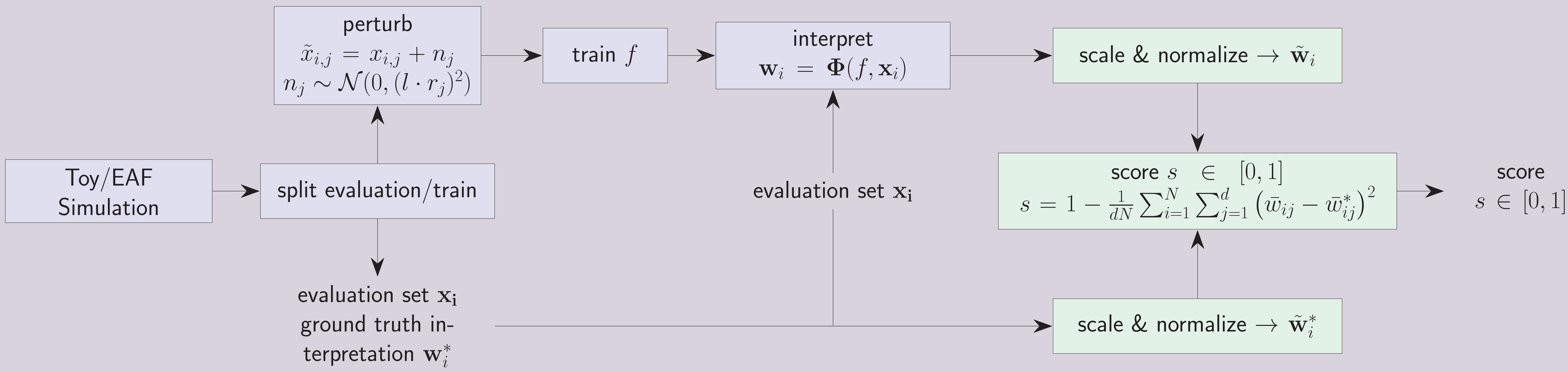
- Problem: eXplainable Artificial Intelligence (XAI) methods not evaluated for performance in noisy settings
- Approach: evaluation pipeline, including simulated dataset generation and comparing explanations to ground truth effects
- Results: Explainer performance directly tied to model performance, 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
- Robustness and correctness are not quantified – need to evaluate noise robustness & correctness of XAI in averse situations
- Ground truth effect $\mathbf{w}_i^*$ not available in real-world data → **simulated datasets with ground truth**!
- Scoring for XAI methods difficult → **evaluate using custom methods!**
- Different kinds of XAI methods
  - *effects*: Gradient, SG, ALE-kNN
  - *attribution*: LIME, SHAP

Our Evaluation methodology

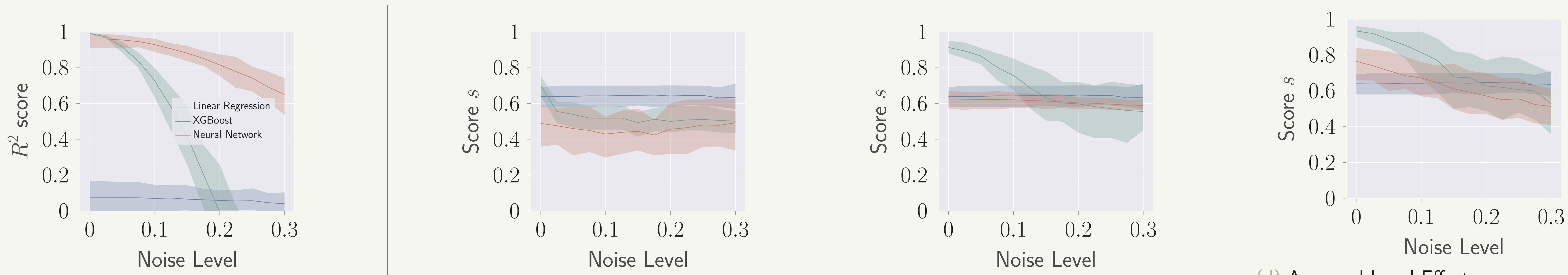
- Solve scaling & alignment issues
- Artififcially perturb dataset using noise $n_j \sim \mathcal{N}(0, (l \cdot r_j)^2)$ based on data range r_j
- Train model $f(\mathbf{x})$
- Infer local interpretations $\mathbf{w}_i = \mathbf{\Phi}(f, \mathbf{x}_i)$
- Calculate score $s \in [0, 1]$



Results

- Toy dataset: polynomial generator
  - Generate 1000 samples
  - Calculate ground truth $\mathbf{w}^*$ using automatic differentiation

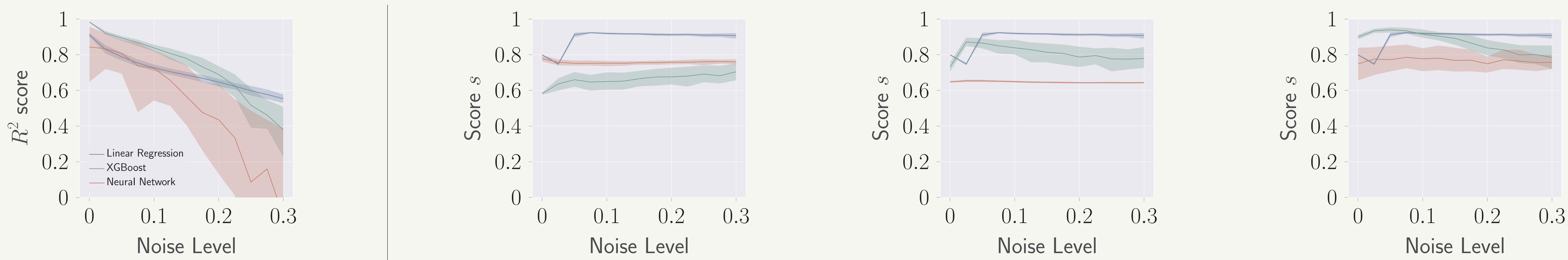
Figure: Score s on toy data with varying levels of noise on the different combinations of explainers and Machine Learning (ML) models.



(a) R2 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 whole simulation

Figure: Score s on EAF data with varying levels of noise on the different combinations of explainers and ML models.



(a) R2 Score (b) Gradient (c) SG (d) ALE-kNN