Robustness of Explainable Artificial Intelligence in Industrial Process Modelling

Benedikt Kantz [1] **Clemens Staudinger** [2] **Christoph Feilmayr** [2] **Johannes Wachlmayr** [3] **Alexander Haberl** [2] **Stefan Schuster** [2] **Franz Pernkopf** [1 4]

Abstract

eXplainable Artificial Intelligence (XAI) aims at providing understandable explanations of black box models. This paper evaluates current XAI methods by scoring them based on ground truth simulations and sensitivity analysis. To this end, we used an Electric Arc Furnace (EAF) model to understand better the limits and robustness characteristics of XAI methods such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), as well as Averaged Local Effects (ALE) or Smooth Gradients (SG) in a highly topical setting. These XAI methods were applied to various black-box models and then scored based on their correctness compared to the ground-truth sensitivity of the data-generating processes using a novel scoring evaluation methodology over a range of simulated additive noise. The resulting evaluation shows that the capability of the Machine Learning (ML) models to capture the process accurately is, indeed, coupled with the correctness of the explainability of the underlying data-generating process. We show the differences between XAI methods in their ability to predict the true sensitivity of the modeled industrial process correctly.

1. Introduction

ML approaches have the power to model complex dependencies in demanding tasks such as industrial processes. However, the behavior of these industrial processes that rely on complex, non-linear interactions often needs to be fully understood. This results in the need for algorithms to understand and interpret how these ML models arrive at specific predictions and how they might react to certain perturbances in the input. In the last years, there has been an effort to provide explanations to the ML model predictions using XAI (Lundberg & Lee, 2017; Ribeiro et al., 2018; Alvarez-Melis & Jaakkola, 2018; Shrikumar et al., 2017).

Most of these works, even if they focus on the robustness and trustworthiness of the XAI method, have the shortcoming that they can only be evaluated through surrogate measures (Crabbé & van der Schaar, 2023), and the ground truth sensitivity of the evaluated datasets cannot be appropriately calculated (Alvarez-Melis & Jaakkola, 2018). Some existing approaches rather use data augmentation (Sun et al., 2020) or create measures estimating the importance of the features (Yeh et al., 2019); further related work is provided in Section A.3. None of these systems, to the best of our knowledge, consider the ground truth sensitivity, or gradient, of the data-generating process that created the dataset. However, modeling the sensitivity to the inputs is vital to understanding the underlying process using proxy ML models learned solely on data.

In this paper, we introduce data-driven evaluation of different XAI methods using a simulated process of an EAF model and its ground truth sensitivity, providing insights into the actual limits and robustness properties of state-of-the-art ML models and interpretability approaches. We propose to use a specifically generated dataset and perform perturbations to analyze this robustness empirically. Two central problems, however, arose when scoring these XAI methods and comparing them to a ground truth sensitivity:

1. The feature importance scores over the feature dimensions are not within the same magnitude and range, requiring scaling (Shrikumar et al., 2017).
2. The relative sizes of different XAI feature importance scores are not necessarily aligned to each other (Lundberg & Lee, 2017; Apley & Zhu, 2019), requiring normalization.

Therefore, we introduce a novel evaluation methodology to solve these problems. We analyze how well the XAI methods explain the sensitivity of the input features, compared

[1]Signal Processing and Speech Communication Laboratory, Technical University Graz, Graz, Austria [2]voestalpine Stahl GmbH, Linz, Austria [3]K1-MET GmbH, Linz, Austria [4]Christian Doppler Laboratory for Dependable Intelligent Systems in Harsh Environments, Graz, Austria. Correspondence to: Benedikt Kantz <benedikt.kantz@tugraz.at>, Franz Pernkopf <pernkopf@tugraz.at>.

Accepted at the 1st Machine Learning for Life and Material Sciences Workshop at ICML 2024. *Updated version 2024-09-26: changed models and rerun of experiments; improved stability of results.*

to a known ground truth sensitivity.

2. Data-generating processes

This section outlines the two data generation processes: a toy dataset and the EAF process simulation, including the process variables. These specifically generated datasets are necessary, as other datasets do not provide the ground truth effects $\mathbf{w}_i^*$ of the functions at the datapoints $\mathbf{x}^i$.

2.1. Toy dataset

To prove the effectiveness of our evaluation methodology, we first build a small polynomial data-generating system of the form

$$f(x_1, x_2) = k_1 x_1^2 + k_2 x_2^2 + k_3 x_1 x_2 + k_4 x_1 + k_5 x_2 + k_6 \quad (1)$$

with the coefficients k_d being drawn from $k_d \sim \mathrm{Uniform}(0, 1)$ once. The function effects, or gradients, were estimated using automatic gradient calculation using PyTorch (Paszke et al., 2019). This data-generating process was used to generate 5000 samples $f(x_1, x_2)$, where $x_1, x_2 \sim \mathrm{Uniform}(-5, 5)$, which were utilized for the evaluation process.

2.2. EAF process

The choice of an EAF model as the data-generating process was owed to various factors. First, the EAF process is relevant in the steel industry as it promises, given enough clean energy, greenhouse gas emission reduction compared to traditional steel production in blast furnaces (De Ras et al., 2019). Furthermore, EAFs have been studied and modeled for a long time using many different approaches and modeling strategies, from their electrical characteristics (Billings et al., 1979; Boulet et al., 2003) to their chemical processes and internal interactions (Zhang & Fruehan, 1995; Basu et al., 2008). These works helped build an accurate chemical simulation of an EAF, representing the real-world system well while keeping it simulatable and manageable in parameter space (Binti Ahmad Dzulfakhar et al., 2023). The functional complexity is sophisticated, providing an interesting process to be modeled by ML models.

Furthermore, the data generating process is non-independant and identically distrubuted (iid.), as the EAF model has timesteps, and, when considering these, dependencies between each simulated furnace tapping arise, making modeling even more challenging. The model itself simulated the individual zones of the EAF reactor as homogeneous zones, with the same temperature and uniform mixture across the whole zone. These zones are the gas zone, solid metal zone, liquid slag zone, liquid metal, and solid slag zone. Additionally, the reactor is modeled as a few discrete parts, particularly the roof and walls. The liquid metal and slag zones are the ones where measurements were simulated during tapping.

The EAF model (Binti Ahmad Dzulfakhar et al., 2023) was converted into a Python module, where automatic differentiation tools (Paszke et al., 2019) were used to generate ground truth sensitivities of the input parameters. This simulated experiment was repeated over sampled combinations of a subset of different input parameters, which were the oxygen lance rate, oxygen for post-combustion, power of the arc, carbon injection rate, ferromanganese injection rate, and the mass addition rate of solids. Furthermore, auxiliary properties within the simulated tapped material were recorded, specifically the ratios of silicon dioxide and iron oxide in the slag and the temperature of the liquid slag and metal. The observed target variable, the ratio of carbon in the tapped steel from the furnace, was recorded, too. Each tapping was considered as one data sample $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, consisting of the observed variables. The gradient of the input parameters concerning the output was also calculated and accumulated at the simulation's timesteps. This simulation resulted in a dataset of about $n \approx 10^4$ samples after the removal of numerical outliers due to instabilities after simulating the furnace for about a week. The process was furthermore restarted after each parameter change.

2.3. Perturbing the Dataset

While the datasets at hand provided a perfect ground truth sensitivity for interpretations, there was still no proper way to assess the robustness of the ML models in combination with the XAI methods. To this end, the dataset was artificially perturbed using noise in two ways. The source of the noise was Gaussian noise added to the feature j of sample i, except the target variable, using

$$\tilde{x}_{i,j} = x_{i,j} + n_j \quad (2)$$

where $n_j \sim \mathcal{N}(0, l \cdot (\max_{i \in 1,\ldots,n} x_{i,j} - \min_{i \in 1,\ldots,n} x_{i,j}))$, $x_{i,j}$ being the j-th feature of the observation $\mathbf{x}_i = \{x_{i,0}, x_{i,1}, \ldots, x_{i,d}\}$. The 0-th feature is the target $y_i \in \mathcal{Y}$. $l \in [0, 1)$ is the selected noise level of the experiment.

3. Scoring Methodology for Local Explanations

We developed an evaluation methodology (see Figure 1) to quantify the effect of the perturbances on the ML models and the interpretation, as there is no such common measure to cater to heterogeneous types of feature importance and interpretability measures, as discussed in Section A.3. Furthermore, the problems that make comparisons difficult, as mentioned in Section 1, were addressed using this scoring methodology.

We denote the output of each XAI method as $\mathbf{w}_i =$

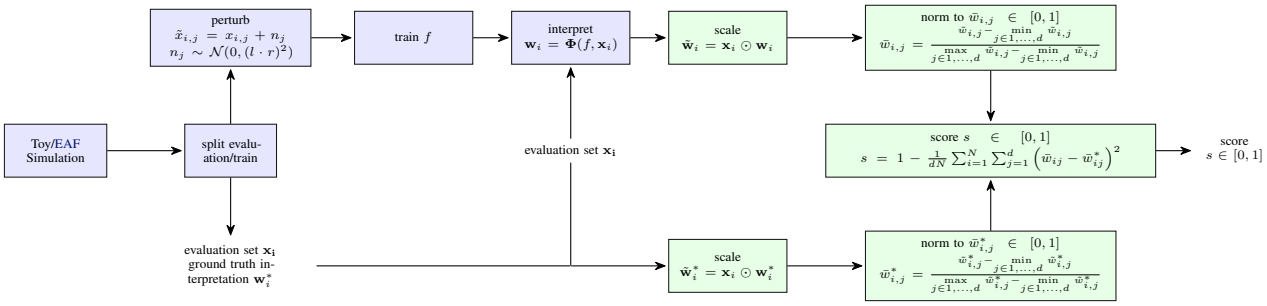


Figure 1: Evaluation methodology, see text for further details and choices of measures.

$w_{i,1}, \dots, w_{i,d}$, independent of the underlying explainer. All ground truth sensitivity values from the data-generating processes are denoted as $\mathbf{w}_i^*$. All input features of the data-generating functions were used in this evaluation system. To solve the first problem outlined in the introduction, the feature scores $w_{i,j}$ were scaled using the features values themselves, using

$$\tilde{w}_{i,j} = w_{i,j} \cdot x_{i,j}. \tag{3}$$

This scaling helps to adjust the magnitude of the XAI method results to the appropriate scale. The second problem is approached by min-max normalization on each data sample to indicate their relative strength within a sample. The scaling results in an indication of the weights' approximate strength by scaling them to a $[0, 1]$ range using

$$\bar{w}_{i,j} = \frac{\tilde{w}_{i,j} - \min\limits_{j \in 1,\dots,d} \tilde{w}_{i,j}}{\max\limits_{j \in 1,\dots,d} \tilde{w}_{i,j} - \min\limits_{j \in 1,\dots,d} \tilde{w}_{i,j}}. \tag{4}$$

Both scaling operations are performed for the ground truth derivatives $w_{i,j}^*$ of the data-generating process, too, leading to the normalized ground truth $\bar{w}_{i,j}^*$. Using both results, the score of the local interpretation can finally be computed using the Brier Score (Brier, 1950), a probabilistic scoring method initially intended for weather forecasts. It is, in essence, the Mean Squared Error (MSE), and the score s_i for one observation is thus calculated using

$$s_i = \frac{1}{d} \sum_{j=1}^{d} \left(\bar{w}_{i,j} - \bar{w}_{i,j}^*\right)^2. \tag{5}$$

The final score s for one set of n observations is computed by averaging each samples score s_i using

$$s = 1 - \frac{1}{n} \sum_{i=1}^{n} s_i. \tag{6}$$

This averaging should return a score of $s = 1$ if the interpretations completely align with the ground truth and lower if there are discrepancies. This scoring methodology, due to the comparison to the ground truth effect, focuses, therefore, not on the feature's importance but rather on the correctness of the effects that a feature has at a certain data point.

3.1. Evaluation Process

The evaluation methodology is performed 50 times over a randomly sampled fold of 10 percent of the samples from the datasets. However, due to the non-iid data of the EAF, care is taken during sampling. Therefore, the simulation runs are sampled so that no data from one run can be taken into the evaluation and training set. The sampled training set (90% of the samples) is then perturbed using a range of different noise levels using the approaches from Section 2.3. Figure 1 illustrates the whole evaluation methodology, showing how the explanations are scaled and used throughout the process. The blue part of the graph illustrates existing work and systems, while the green part shows the novel scaling and scoring scheme.

3.1.1. Black-box Models

The regressor ML models $f(\mathbf{x}_i)$ used for this evaluation are a linear regression model, a neural network ensemble consisting of four networks with three layers of 32 neurons with Rectified Linear Unit (ReLU) activation modeled in PyTorch (Paszke et al., 2019), and the Light Gradient Boosting Machine (LightGBM) regressor (Ke et al., 2017). Explanations $\mathbf{w}_i = \mathbf{\Phi}(f, \mathbf{x}_i)$ are generated using the XAI methods discussed in the next section. The necessary gradients are calculated using either the parameters directly from linear regression, automatic differentiation for the neural network, or finite differences for the LightGBM tree. These local explanations are only generated on the unperturbed and complete evaluation set. This scheme allows us to test how well and accurately ML models can learn feature importances even in noisy settings.

3.1.2. Explanation Methods

The introduced score s and the three ML models $f(\mathbf{x})$ are then used in five different XAI methods. An overview of the general XAI landscape is provided in Section A.1, showing the different types of XAI methods and why we chose local explanations. The selected local XAI methods can generally be categorized in either *Effect-based Methods* (EM) and *Additive Methods* (AM). The first XAI method describes the effects the input shift has on the output; the latter how much the feature contributes to the output - usually in a sparse form. The following XAI methods are used (theoretical details in Section A.2):

- The gradient baseline simply takes the gradient of the input of the ML model - a very simple approach (EM).
- Improving upon that, the SG method averages the gradient over $k = 10$ neighbors sampled using k-Nearest Neighbors (kNN) (Yeh et al., 2019) (EM).
- Next, the ALE-kNN (Apley & Zhu, 2019) uses a simple Gaussian conditional distribution with a fixed $\sigma^2 = 0.2$ scaled by the feature range, a resolution of $n_{samples} = 50$ bins over the feature range and $k = 10$ samples for the kNN selection (EM).
- LIME (Ribeiro et al., 2016) uses default values, with no modifications (AM).
- The SHAP (Lundberg & Lee, 2017) methods are used with default parameters. Different SHAP methods, however, are used, depending on the ML model - Tree SHAP for LightGBM and Gradient SHAP for the neural network ensemble and Gaussian Process (GP) (AM).

4. Results & Discussion

This section presents the evaluation results, beginning with an overview of how noise affects the different XAI approaches in combination with different ML models. Sanity checks with just noise are also performed, providing a lower empirical bar for s of the evaluated systems.

4.1. Robustness Results

The core results of this paper from both datasets are shown in Figure 2 and 3, where we show the effect of progressively applying more noise to the training data. Various black-box models are then fitted to this perturbed training set, and XAI methods are evaluated using the scoring method proposed in Section 3.

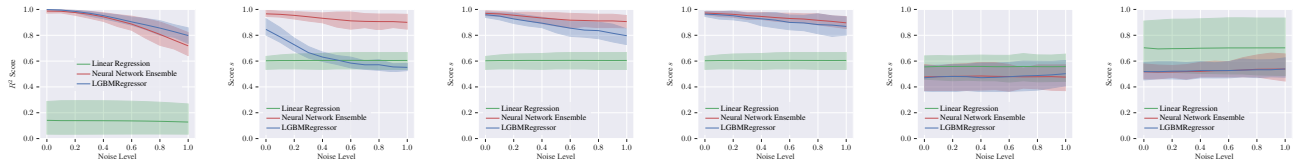
The first graphs of either dataset, Figures 2a and 3a, show the R^2 score, a metric to evaluate regression problems. The score can achieve a maximum of 1 if the predictions are perfect, 0 if they predict the mean, and arbitrarily negative if the prediction is worse than the mean (Chicco et al., 2021). These scores show that the increasing noise has, as expected, adverse effects on the performance of the evaluation set. The following graphs of the toy dataset, Figures 2b through 2d show that the effect-based XAI methods are strongly dependent on the performance of the trained ML models. Linear regression fails to capture most relations while LightGBM and the neural network work quite well, especially in regimes of low noise using a robust explainer like ALE-kNN. The additive methods, shown in Figures 2e and 2f, are not able to capture the sensitivity present in the ground truth $\mathbf{w}^*$ even without noise.

Similarly, the EAF results show that the explainer performance recorded is coupled with the ML model performance, as the effect-based methods fall with increasing noise. This rising noise plays, again, a vital role in the lowering of the scores of the LightGBM for the more robust XAI methods, SG and ALE-kNN, as seen in Figures 3c and 3d. These two and the raw gradient of Figure 3b show two further interesting findings: first, when the LightGBM performance is good, the explainer score of the LightGBM is relatively constant over the range of noise levels. Second, the LightGBM needs some initial noise on this dataset to start modeling the relations well, which can be observed for all effect-based methods, where all XAI method's scores increase when adding an initial bit of noise. This rise could be due to a regularizing effect of the noise on the gradients, effectively creating more truthful averages of gradients when adding noise. The additive methods are again not as effective in explaining the ground truth sensitivity as evident in Figures 3f and 3e

4.2. Random baseline: Empirical lower bounds for Results

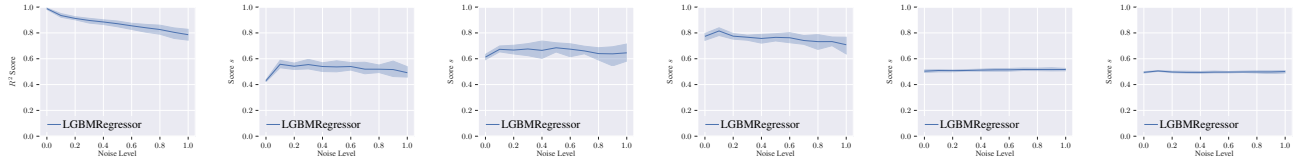
We additionally empirically calculate the random lower bounds using noisy data to ensure the evaluation measures worked. This check was performed by training the ML model on the correct, slightly perturbed data using a noise level of $l = 0.5$. The ML model was then given random evaluation data, on which the scores for the interpreters were calculated. The evaluation dataset was set to $x_{ij} \sim \mathcal{N}(\mu_j, \sigma_j)$, where μ_j and σ_j are the mean and standard deviation of the training dataset; the used dataset is the toy dataset. The results for this perturbation can be seen in Table 1, where all XAI methods approached a score of about 0.5 to 0.6, indicating that this is indeed similar to the worst results observed above.

In the next experiment, we trained the ML model on completely noisy data, again with $x_{ij} \sim \mathcal{N}(\mu_j, \sigma_j)$, but scored on the correct evaluation set, shown in Table 2. This produced, expectedly, even similar performance metrics. This indicates that the random baseline for this scoring method is around $s = 0.55$. The neural network ensemble, however, managed to retain some performance with the effect-based



(a) (b) (c) (d) (e) (f)

Figure 2: Score s on toy data with varying levels of noise on the different combinations of explainers and ML models. The shaded area is the 90th and 10th percentile over 50 experiments with random sampling. (a) R^2 score, (b) Gradient score s, (c) SG score s, (d) ALE-kNN score s, (e) LIME score s, and (f) SHAP score s.



(a) (b) (c) (d) (e) (f)

Figure 3: Score s on EAF data with varying levels of noise on the different combinations of explainers and the LightGBM. The shaded area is, again, the 90th and 10th percentile over 50 experiments with random sampling. (a) R^2 score, (b) Gradient score s, (c) SG score s, (d) ALE-kNN score s, (f) LIME score s, and (e) SHAP score s,

explainers around $s = 0.75$ with a high variability.

4.3. Discussion

This investigation of the robustness of XAI methods showed that the noise and the predictive performance influence these approaches. This is especially true for gradient-based XAI approaches. The SG as well as the cohort-based ALE-kNN works well. The properties of the custom cohort approach foster the correctness of the interpretations, as there is an uplift in performance, especially in combination with the tree-based LightGBM system. However, both LIME and SHAP are in the category of additive methods. This weakness makes them unsuitable for interpretations that call for scores reflecting the input sensitivity on the output, i.e., reflecting the feature *effect* of the output.

The low performance of the additive feature importance scores can also be partially attributed to the metric preferring the effect-based scoring methods, as the related work by Yeh et al. (2019) notes. Notwithstanding, we showed that effect-based scoring methods are highly dependent on the performance of the ML model to accurately reflect the ground truth importance scores w_{ij}^* and that a single gradient of the ML model is often not enough to estimate them correctly.

5. Conclusion & Future Work

We showed how different XAI methods are affected based on the predictive performance of the ML models. This work focused on model-agnostic post-hoc explanations for local data samples, as these could be evaluated using numeric, data-driven approaches. Of these XAI methods, SHAP, LIME, SG, and a local version of ALE were chosen. These were evaluated using a novel evaluation process focused on scaling the feature importance scores to a similar magnitude within one sample, normalizing them to the same range as the ground truth effects, and calculating the distance to the ground truth effects reference. This ground truth was generated using a chemical simulation of an EAF model, providing the necessary ground truth sensitivity for comparison and evaluation. This data-generating distribution was chosen based on the maturity of the EAF models for these real-world processes and current interest in the technology due to its promise of cleaner steel production. A toy example was also initially used to test the approaches on a limited and known nonlinear dataset.

Noise analysis over a range of perturbances of the initial dataset was performed using this evaluation methodology. The resulting analysis concludes that LightGBM, in combination with smooth gradient-based XAI methods, can approximate both the target values and the ground truth interpretations very well, even in noisy environments. LIME and SHAP were not successfull in correctly finding the ground truth feature importance scores, probably due to their differing approaches to interpreting the feature importance scores. Furthermore, some of these XAI methods and ML models showed a higher variance, indicating that they varied between sampling runs and were affected by the high noise. Sanity checks on the validity of the evaluation process were carried out using noise, first, as evaluation data, and then as training data. Both tests showed that the scores

Table 1: Results of sanity check with Gaussian noise as evaluation inputs on the toy dataset (± one standard deviation).

MODEL	s_{XAI} GRAD	s_{XAI} SG	s_{XAI} ALE-KNN	s_{XAI} LIME	s_{XAI} SHAP
LGBMREGRESSOR	0.50±0.02	0.52±0.04	0.54±0.04	0.49±0.02	0.50±0.03
LINEAR REGRESSION	0.57±0.03	0.57±0.03	0.57±0.03	0.51±0.02	0.49±0.03
NEURAL NETWORK ENSEMBLE	0.52±0.03	0.53±0.03	0.54±0.04	0.50±0.02	0.50±0.03

Table 2: Results of sanity check using Gaussian noise as training inputs on the toy dataset (± one standard deviation).

MODEL	s_{XAI} GRAD	s_{XAI} SG	s_{XAI} ALE-KNN	s_{XAI} LIME	s_{XAI} SHAP
LGBMREGRESSOR	0.50±0.02	0.48±0.07	0.48±0.13	0.49±0.10	0.50±0.04
LINEAR REGRESSION	0.51±0.10	0.51±0.10	0.51±0.10	0.47±0.16	0.45±0.24
NEURAL NETWORK ENSEMBLE	0.74±0.16	0.75±0.17	0.75±0.17	0.49±0.11	0.47±0.11

are around $s = 0.55$ since the ML models cannot learn the importance at all.

5.1. Future Work

There is an apparent need to quantify the uncertainty of the XAI methods, as this high variability of the feature importance scores of ML models with high uncertainty in noisy environments distorted the feature interpretation significantly. There are already works investigating such uncertainties for XAI (Löfström et al., 2024; Zhao et al., 2021; Slack et al., 2021), however, none of these address the effect-based approaches where the feature importance score reflects the change of the output with respect to the input.

The empirical evaluation of feature importance scores, especially from LIME and SHAP, could also be further investigated by comparing different metrics on a ground truth dataset. Further improvements on the measures of infidelity and sensitivity (Yeh et al., 2019), combined with the consideration of a known ground truth feature importance and deeper analysis of noise could also lead to further understanding of robustness and failure cases of XAI.

Acknowledgements

The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. We furthermore thankfully acknowledge the financial support of the project by voestalpine Stahl GmbH. The authors gratefully acknowledge the funding support of K1-MET GmbH, whose research program is supported by COMET (Competence Center for Excellent Technologies), the Austrian program for competence centers. COMET is funded by the Austrian ministries BMK and BMDW, the Federal States of Upper Austria, Tyrol, and Styria, and the Styrian Business Promotion Agency (SFG).

References

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. URL http://arxiv.org/abs/1806.08049.

Apley, D. W. and Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models, 2019.

Basu, S., Lahiri, A. K., and Seetharaman, S. Activity of iron oxide in steelmaking slag. *Metallurgical and Materials Transactions B*, 39(3):447–456, May 2008. ISSN 1543-1916. doi: 10.1007/s11663-008-9148-4. URL http://dx.doi.org/10.1007/s11663-008-9148-4.

Bhatt, U., Weller, A., and Moura, J. M. F. Evaluating and aggregating feature-based model explanations, 2020.

Billings, S., Boland, F., and Nicholson, H. Electric arc furnace modelling and control. *Automatica*, 15(2):137–148, 1979. ISSN 0005-1098. doi: https://doi.org/10.1016/0005-1098(79)90065-7. URL https://www.sciencedirect.com/science/article/pii/0005109879900657.

Binti Ahmad Dzulfakhar, A. S., Charaschanya, T., Charos, A., Duru, N., Khan, Fatima Elzahra Lee, S., Sim, E. Y., Tonus, A., Walker, P., and Woolley, R. Eaf modelling: Process design of key reactor unit, November 2023. URL https://github.com/SayYoungMan/EAF_modelling.

Boulet, B., Lalli, G., and Ajersch, M. Modeling and control of an electric arc furnace. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 4, pp. 3060–3064 vol.4, 2003. doi: 10.1109/ACC.2003.1243998.

Brier, G. W. Verification of forecasts in terms of probability. *Monthly Weather Review*, 78(1):1–3, January 1950. ISSN 1520-0493. doi: 10.1175/1520-0493(1950)078⟨0001:vofeit⟩2.0.co;2. URL http://dx.doi.org/

10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Chicco, D., Warrens, M. J., and Jurman, G. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, July 2021. ISSN 2376-5992. doi: 10.7717/peerj-cs.623. URL http://dx.doi.org/10.7717/peerj-cs.623.

Crabbé, J. and van der Schaar, M. Evaluating the robustness of interpretability methods through explanation invariance and equivariance, 2023.

De Ras, K., Van de Vijver, R., Galvita, V. V., Marin, G. B., and Van Geem, K. M. Carbon capture and utilization in the steel industry: challenges and opportunities for chemical engineering. *Current Opinion in Chemical Engineering*, 26:81–87, 2019. ISSN 2211-3398. doi: https://doi.org/10.1016/j.coche.2019.09.001. URL https://www.sciencedirect.com/science/article/pii/S221133981930036X.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning, 2017.

Huang, W., Zhao, X., Jin, G., and Huang, X. Safari: Versatile and efficient evaluations for robustness of interpretability, 2023.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.

Löfström, H., Löfström, T., Johansson, U., and Sönströd, C. Calibrated explanations: With uncertainty information and counterfactuals. *Expert Systems with Applications*, 246:123154, July 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.123154. URL http://dx.doi.org/10.1016/j.eswa.2024.123154.

Munn, M. and Pitman, D. *Explainable AI for practitioners*. O'Reilly Media, Sebastopol, CA, November 2022.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900654116.

Naser, M. An engineer's guide to explainable artificial intelligence and interpretable machine learning: Navigating causality, forced goodness, and the false perception of inference. *Automation in Construction*, 129:103821, September 2021. ISSN 0926-5805. doi: 10.1016/j.autcon.2021.103821. URL http://dx.doi.org/10.1016/j.autcon.2021.103821.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017. URL http://arxiv.org/abs/1704.02685.

Slack, D. Z., Hilgard, S., Singh, S., and Lakkaraju, H. Reliable post hoc explanations: Modeling uncertainty in explainability. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=YUEFlzlG_0c.

Sun, Y., Chockler, H., Huang, X., and Kroening, D. *Explaining Image Classifiers Using Statistical Fault Localization*, pp. 391–406. Springer International Publishing, 2020. ISBN 9783030586041. doi: 10.1007/978-3-030-58604-1_24. URL http://dx.doi.org/10.1007/978-3-030-58604-1_24.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., and Ravikumar, P. On the (in)fidelity and sensitivity for explanations, 2019.

Zhang, Y. and Fruehan, R. J. Effect of gas type and pressure on slag foaming. *Metallurgical and Materials Transactions B*, 26(5):1088–1091, October 1995. ISSN 1543-1916. doi: 10.1007/bf02654112. URL http://dx.doi.org/10.1007/BF02654112.

Zhao, X., Huang, W., Huang, X., Robu, V., and Flynn, D. Baylime: Bayesian local interpretable model-agnostic explanations, 2021.

A. Background & Related Work

A.1. Interpretable and Explainable ML

Before diving into the XAI methods themselves, the notion of interpretability and explainability of ML models has to be examined first. These terms span a wide field of approaches, with the common goal of enabling the end users or creators of a system to foster their understanding on *why*, and sometimes *how* ML models arrive at certain conclusions. However, the presentation of these explanations can vary widely, from visual explainers to mathematical formulas. This paper will focus on the latter, as we will take advantage of that they are easier to compare and evaluate numerically in a data-driven environment. Most methods can, despite this variety in modalities, generally be categorized into either *interpretable systems* or so-called *post-hoc* methods (Murdoch et al., 2019; Doshi-Velez & Kim, 2017).

A.1.1. Interpretable Systems

ML models can achieve inherent interpretability through different modalities, which, however, share the constraint of understandable processes that can be comprehended by a single person through a reasonable timeframe and effort. This limitation, coined by Murdoch et al. (2019) as "simulatability", greatly inhibits both the choice of system and the function space of the considered models. This leads, depending on the reduction in expressive power, to a loss of predictive performance and has to be considered before choosing such a model. The (numerical) evaluation of these systems is furthermore hindered by the wide-ranging type of approaches and often depends on human evaluation of these systems (Murdoch et al., 2019).

One of the two common ways to increase the range of available systems again can be the introduction of sparsity, where there is an initial, more complex function and feature space that can later be reduced by fitting the model to the data and reducing the set of expressions needed for decisionmaking. Similarly, the model could be broken up into smaller modules, where each is very well understood on its own - a popular choice for these modularized, linear components would be ensemble-based approaches with only a few models (Murdoch et al., 2019).

A.1.2. Post-hoc Interpretability

These interpretable approaches, however, do not foster the understanding of the model if there are no alternatives to the current model due to performance limitations or other constraining factors limiting choice. In these cases, post-hoc approaches can offer interpretability for black-box models that would be too complicated for humans to comprehend easily. These can be further categorized into global (dataset-level) and local (instance-level) approaches to interpretability. Global explanations are of interest for an overview of how the model considers certain features in general or for statistical analysis. If the interest lies in specific examples, or even failure cases such as adversarial examples, where a prediction can be maliciously flipped, a local explanation is often the needed approach (Murdoch et al., 2019). Furthermore, one can also group or cluster features and analyze these cohorts with global methods, achieving a similar overview-level interpretation on a smaller neighborhood (Munn & Pitman, 2022). All levels utilize similar techniques to arrive at solutions explaining the decision, with the core concept being called *feature scores*. These methods measure how influential a certain feature is, either on the local or global level, and assign it a score based on the perturbances it might achieve. These methods such as *partial dependence plots (PDPs)* and *ALE* (Apley & Zhu, 2019) (in the 1D case), or their local counterpart, *SG*, describe how varying a single feature affects the output, with the assumption that features of interest are not correlated. We will call these methods *effect-based* methods throughout this paper. *Feature interactions*, on the other hand, analyze exactly these contributions: how two or more features influence the output in conjunction (Naser, 2021). Other methods, such as *LIME* utilize simple, local representations of models as surrogates to better explain the behavior of models in certain neighborhoods (Ribeiro et al., 2016; Munn & Pitman, 2022). Both LIME and SHAP, another explanation methodology, form a group of systems we will call *feature importance scores*, where each feature is assigned a certain contribution to the output.

A.2. Chosen Approaches

Of these methods, the local and cohort-based post-hoc approaches for feature importance are of interest for the evaluation task at hand, as these allow for comparison to known feature influences from the simulated data-generating process. This is why we chose ALE on a small cohort chosen using kNN (ALE-kNN), SG, LIME and SHAP.

A.2.1. ALE-kNN AND SG

The intuition behind ALE (Apley & Zhu, 2019) is to simply accumulate the local effects (derivatives) of a function (model) weighted by a certain prior. This prior is the main difference to PDPs, as these do not consider the conditional distribution $p_{2|1}(x_2|x_1)$ of the sample, just the marginal over the selected variable. This results in a very intuitive and straightforward formulation of the ALE effect $f_{1,ALE}(x_1)$ on x_1 for a two-dimensional function ($d = 2$, $f(x_1, x_2)$) as

$$f_{1,ALE}(x_1) = \int_{x_{min,1}}^{x_1} \int p_{2|1}(x_2|z_1) f^1(z_1, x_2) dx_2 dz_1 - \text{constant} \tag{7}$$

where $f^1(x_1, x_2)$ is the first derivative with respect to the variable of interest ($\frac{\partial f(x_1,x_2)}{\partial x_1}$). $x_{min,1}$ is chosen such that it just touches the lower support of the conditional distribution, approximating the integral over the whole distribution. The first integration variable x_2 integrates over all samples of the second, unrelated feature (or more in real-world examples), while z_1 integrates the effect over the conditional distribution. The resulting uncentered ALE is retrieved when ignoring the constant, which can be centered by subtracting the average effect. This method is usually applied on a dataset level and generates visual plots, showing the influence of variables over the whole range of the input (Apley & Zhu, 2019).

However, to apply this method to a local data sample, the k nearest neighbors of the sample are chosen, and the resulting local cohort (Munn & Pitman, 2022) is used as the basis for the uncentered ALE of this sample. This local cohort version of the ALE is very closely related to the SG method, where the gradients or, more general, the explanations are averaged over a small neighborhood (Yeh et al., 2019). The version using ALEs has, however, the advantage of the introduction of the conditional distribution, effectively weighting the effect locally.

A.2.2. LIME

LIME (Ribeiro et al., 2016), offers feature importance scores through local surrogate models, learning a sparse representation of the neighborhood. This approach combines almost all concepts outlined in the overview: simulatability, sparsity, local cohorts, and surrogate models. The resulting method returns feature importance scores. The algorithm itself consists of three major steps: sample selection, feature importance calculation, and a final selection process called "submodular pick". For local explanations, only the first two steps are necessary.

Sample selection searches a training dataset for similar datapoints using a configurable distance metric, and evaluates the model for each. This small set is then used for fitting the surrogate model to the local neighborhood. A linear regressor is usually used as a local surrogate. The weights of the resulting model are then directly utilized as the feature importance scores (Ribeiro et al., 2016).

A.2.3. SHAP

The SHAP (Lundberg & Lee, 2017) framework unifies multiple methods for calculating SHAP values, a measure of feature importance. They are based on Shapley values, the only set of values satisfying the three desirable properties for additive feature attribution. This scoring method can be described as a simple linear model of the form

$$g(z') = \phi_0 + \sum_{i=i}^{M} \phi_i z'_i, \tag{8}$$

where the model is similar to the function at $g(z') \approx f(h_x(z'))$ and $z' \in \{0, 1\}^M$ being a binary representation of x', a *simplified input*, using the mapping function $h_x(z') = x$. The parameters ϕ_i describe the actual feature attribution.

While LIME follows the same structure, it does not necessarily fulfill the needed theoretical properties. Lundberg & Lee (2017) introduce, based on this notion, three properties. These properties are

1. Local accuracy: the linear model matches the complex model in the local point;
2. Missingness: the instances with $x'_i = 0$ have no impact, so $\phi_i = 0$; and
3. Consistency: a change of the model output w.r.t. one feature resulting in an output increase should change the attribution ϕ_i only positively.

The exact computation of Shapley values is quite challenging, as it is NP-hard - there are, however, heuristics and simplifications for existing models such as Kernel SHAP, which is model agnostic. Other methods such as Deep SHAP or

Tree SHAP utilize certain model-specific aspects to speed up the computation and provide more exact estimates. All of these approaches are united in the SHAP package (Lundberg & Lee, 2017).

A.3. Related Work

Most of the mentioned methods prove their value either using human evaluation (Ribeiro et al., 2016; Lundberg & Lee, 2017; Ribeiro et al., 2018), with the use of toy examples (Ribeiro et al., 2016; Apley & Zhu, 2019; Shrikumar et al., 2017; Ribeiro et al., 2018) or improvements in theoretical properties (Lundberg & Lee, 2017). None of these works consider robustness to perturbances, unstable outputs, or other chaotic behavior. Some introductions of newer methods, however, are aware of these problems and decide to forego human evaluation in favor of data-driven analysis and robustness studies (Sun et al., 2020). Further works augment this missing robustness consideration in XAI, studying these problems in more depth (Alvarez-Melis & Jaakkola, 2018; Yeh et al., 2019; Bhatt et al., 2020; Huang et al., 2023; Crabbé & van der Schaar, 2023). Alvarez-Melis & Jaakkola (2018) and Yeh et al. (2019) tackle these problems through the maximal local sensitivity, where they search for the maximal change of the explanation to the instance within a neighborhood. This method asses the stability of local explanations and can thus be seen as a measurement for robustness, especially on noisy data. Other methods try to reduce the sensitivity through aggregation by combining different metrics into a more resilient measure (Bhatt et al., 2020). Yeh et al. (2019) introduces, besides the sensitivity of the explanation, the infidelity of explainable approaches, where they measure the squared difference from a projected explanation ($\mathbf{I}^{\mathrm{T}}\Phi(f, \mathbf{x})$) to the difference of the output when subtracting the projection from the input ($f(\mathbf{x}) - f(\mathbf{x} - \mathbf{I})$) using

$$\mathrm{INFD}(\Phi, f, \mathbf{x}) = \mathbb{E}_{\mathbf{I}\sim\mu_{\mathbf{I}}}\left[\left(\mathbf{I}^{\mathrm{T}}\Phi(f, \mathbf{x}) - (f(\mathbf{x}) - f(\mathbf{x} - \mathbf{I}))\right)^2\right] \tag{9}$$

where $\mathbf{I} \in \mathbb{R}^d$ representing the significant perturbances around $\mathbf{x} \in \mathbb{R}^d$ as a random variable, $\mu_{\mathbf{I}}$ the distribution of these perturbances, $f(\mathbf{x})$ the learned model function and $\Phi(f, \mathbf{x})$ the corresponding explaination. While this measure is more data-driven and universal, it still suffers from only using an approximation of the ideal explanation. Furthermore, no studies regarding the robustness with regard to added noise or known noise levels have been performed.