Explainable Artificial Intelligence Beyond Feature Attributions: The Validity and Reliability of Feature Selection Explanations

Raphael Wallsberger[1,*,†], Ricardo Knauer[1,*,†] and Stephan Matzka[1]

[1]University of Applied Sciences Berlin, School of Engineering II – Technology and Life, KI-Werkstatt, 12459 Berlin, Germany

Abstract

Explainable artificial intelligence (XAI) offers powerful tools to increase the transparency of opaque machine learning models. In contrast to feature attribution methods, XAI-based feature selections provide practitioners with a simple, but often more easily interpretable subset of a model's most influential features. In this work, we systematically evaluate feature selection explanations based on Shapley effects and Shapley Additive Global importancE values (SAGE values) across different machine learning algorithms and tabular datasets, and find that they can offer valid and reliable explanations. We derive under which conditions global post-hoc explainers can likely be trusted, laying the groundwork for future research into the validity and reliability of feature selection explanations across a broader range of settings.

Keywords

XAI, Validity, Reliability, Shapley Effects, SAGE.

1. Introduction

Explainable artificial intelligence (XAI) systems have found increasing adoption across industries in recent years. A major driving force has been the call for transparency to not only foster trust among users, but also to comply with regulatory standards [1, 2]. Practitioners frequently use XAI methods to describe a feature's influence on a predictive model via feature attribution or selection, i.e., via assigning a numerical or binary score to each feature. Feature selection approaches are of particular importance in practice because a small subset of the most influential features is often easier to interpret than a list of numerical scores, especially for non-technical users [3].

Out of the variety of feature attribution and selection approaches, Shapley effects [4, 5] or Shapley Additive Global importancE values (SAGE values [6]) have gained increasing popularity for (arguably) three reasons. First, they are based on the Shapley value, a unique solution concept from cooperative game theory that fulfills well-defined desiderata [3, 7]. Second, they

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

†These authors contributed equally.

raphael.wallsberger@htw-berlin.de (R. Wallsberger); ricardo.knauer@htw-berlin.de (R. Knauer); stephan.matzka@htw-berlin.de (S. Matzka)

are model-agnostic, i.e., they can be applied to any predictive model post-hoc. Third, they offer global explanations across the entire dataset, whereas local methods such as SHapley Additive exPlanations (SHAP [8]) exclusively explain single instances. Despite their popularity, there is only limited evidence whether feature selection explanations based on Shapley effects or SAGE values are valid, though, [3, 7], and no evidence whether the selection process is reliable, i.e., whether the selected feature subsets are stable or robust to slight perturbations in the input (via bootstrapping). Intuitively, we expect similar inputs to produce similar explanations. It therefore remains challenging for practitioners to decide when and how these approaches can be trusted and effectively applied.

Our contributions are as follows:

1. **We evaluate feature selection explanations based on Shapley effects and SAGE values** with two common machine learning baselines for small- and medium-sized tabular data: L2-regularized logistic regression and XGBoost [9, 10]. To the best of our knowledge, we are first to assess the selection reliability in addition to the selection validity for these global explanation methods.
2. **We highlight under which conditions Shapley effects and SAGE values can offer valid and reliable explanations**, and show that our conclusions appear to be relatively robust to predictive model choices and input data changes (Sect. 3.2).

2. Related Work

The Shapley value has served as a useful solution concept for feature attribution or selection in XAI. It can be understood as a weighted average over the marginal contributions of a feature to each feature subset. The weights can be axiomatically derived to uniquely define the Shapley value for each feature [3, 7]. Shapley effects approximate this weighted average with respect to the model output [4, 5], SAGE values with respect to the model performance [6]. In terms of the validity, feature selection explanations based on Shapley values are not guaranteed to yield optimal feature subsets [7]. SAGE's global explanation approach, for example, does not necessarily return the best subset, but has been shown to perform better than SHAP's local explanation method for feature selection [7]. Although the reliability of explanations is considered a key open challenge in XAI research [11], its assessment has so far been limited to local explanation approaches such as SHAP [12, 13, 14, 15]. Given that both Shapley effects and SAGE use Monte Carlo simulations to approximate Shapley values, though, the evaluation of their selection reliability is of central importance for transparency and, ultimately, for building trust.

In the next section, we therefore extend the prior research by systematically assessing Shapley effects and SAGE not only in terms of the selection validity, but also in terms of the selection reliability by evaluating how stable or robust these global explanation methods are to small perturbations in the input data.

3. Experiments

In the following, we first provide details on our experimental setup, including the employed datasets and methods as well as evaluation metrics to assess Shapley effects and SAGE in terms of their selection validity and reliability. We then present our experimental results. Overall, we demonstrate that feature selection explanations can be valid and reliable if the number of labels in the smaller class per feature is sufficiently large for a given predictive model.

3.1. Experimental Setup

3.1.1. Datasets and methods

We evaluated Shapley effects and SAGE as feature selectors with two common machine learning baselines for small- and medium-sized tabular data: L2-regularized logistic regression and XGBoost [9, 10], using the PermutationEstimator class.

We expected feature selection with logistic regression to perform reasonably well when the number of labels in the smaller class per feature, or outcome events per variable (EPV), was at least 10 to 15 [16, 17]. Therefore, we leveraged two synthetic binary classification datasets from our prior work - one with 1090 instances, 53 numerical features, and an EPV of about 5; the other with a smaller number of 14 features and thus an EPV of about 18 [18]. Our L2-regularization hyperparameter was tuned using a nested, stratified, 3-fold cross-validation procedure. The number of relevant features per dataset ranged from $k = 2$ to $k = 5$ [18], and we selected the top-k most influential features according to their Shapley effects or SAGE values. As a reference, we compare our feature selection explanations to greedy or optimal feature selection strategies that were recently employed on the same datasets [18].

For XGBoost, we hypothesized that an EPV of at least 200 was needed for valid and reliable selections [19]. We therefore used two different synthetic binary classification datasets from our prior work - one with 20,000 instances, 10 numerical and categorical features, and an EPV of 100 [20]; the other with a decreased class imbalance and thus an EPV of 200. We numerically encoded ordinal features, one-hot encoded nominal features, and used XGBoost with its default settings [20]. The number of relevant features was fixed at $k = 6$ after one-hot encoding, and we again selected the top-k most influential features according to their Shapley effects or SAGE values. To put our feature selection explanations into context, we compare them to explanations based on mean absolute SHAP values as recently investigated in [20].

3.1.2. Evaluation metrics

To assess the selection validity and reliability of our global explanation methods, we used 100 nonparametric bootstrap samples, i.e., 100 random samples from our datasets with replacement. In terms of the validity, we were interested whether the selected feature subsets were true to the data, i.e., how often the top-k most influential features according to Shapley effects or SAGE values matched the k relevant features in our bootstrap samples. With respect to the reliability, we assessed how stable or robust these global explanation methods were to small perturbations in the input (via bootstrapping) with the stability measure proposed by Nogueira et al. (2018) [21]. We regarded stability scores of <0.40 as poor, 0.40 to 0.75 as intermediate to good, and

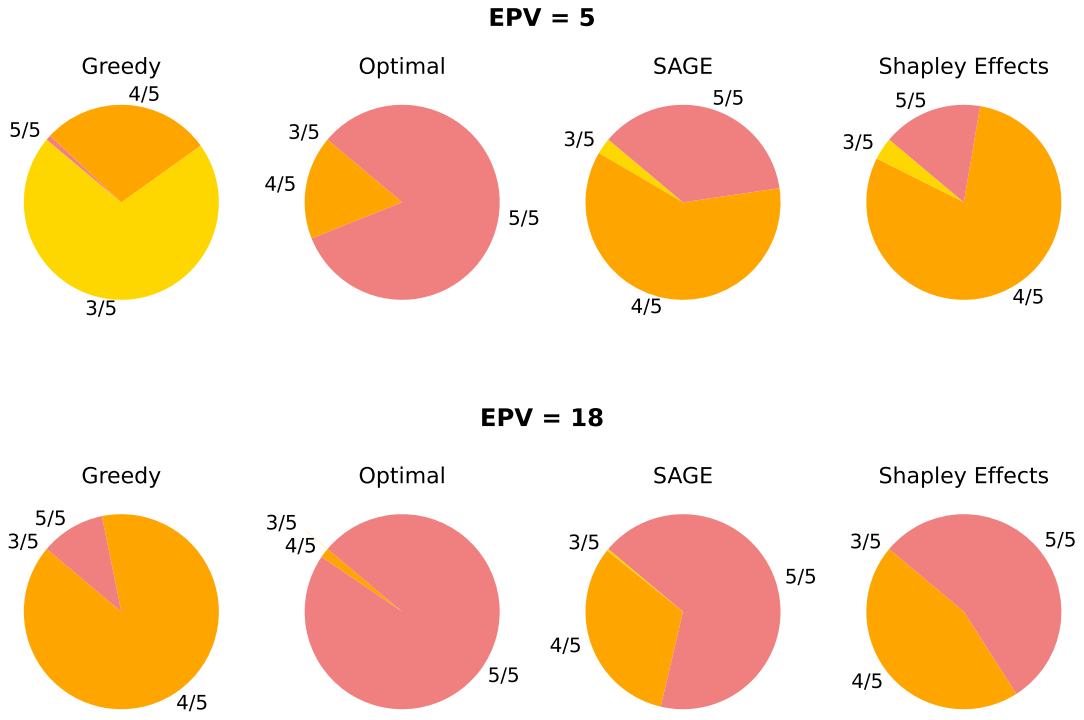

Figure 1: Frequency of correctly selected features for SAGE and Shapley effects, using logistic regression with $k = 5$ at $EPV = 5$ and $EPV = 18$. Greedy and optimal selection strategies for logistic regression have already been used in our prior work and serve as a reference [18]. $5/5$ implies that all informative features have been chosen, i.e., the correct feature subset has been recovered.

>0.75 as excellent [21]. Finally, we evaluated the discriminative performance by computing the mean test area under the receiver operating characteristic curve (AUC) using a (nested) stratified, 3-fold cross-validation procedure [18].

3.2. Experimental Results

Fig. 1 shows the validity for logistic regression, with $k = 5$ as an illustrative example, Fig. 3 the reliability for logistic regression; Fig. 2 and Fig. 4 depict the validity and reliability for XGBoost.

For logistic regression, we observe that both Shapley effects and SAGE perform relatively well, given sufficiently large EPVs. At $EPV = 5$, no selection strategy consistently identifies the correct feature subset across different levels of k. The selection stability is best for greedy selection at 0.88 (95% confidence interval (CI) [0.88, 0.89]). Feature selections based on Shapley effects and best subset selection still yield excellent stabilties at 0.81 (95% CI [0.80, 0.83]) and 0.77 (95% CI [0.75, 0.78]), whereas SAGE achieves the worst stability at 0.74 (95% CI [0.73, 0.76]). At $EPV = 18$, the correct features are most frequently found by optimal selection, followed by SAGE and Shapley effects. With respect to the reliability, the optimal selection strategy performs best with a stability score of 0.94 (95% CI [0.93, 0.96]). Shapley effects and SAGE still

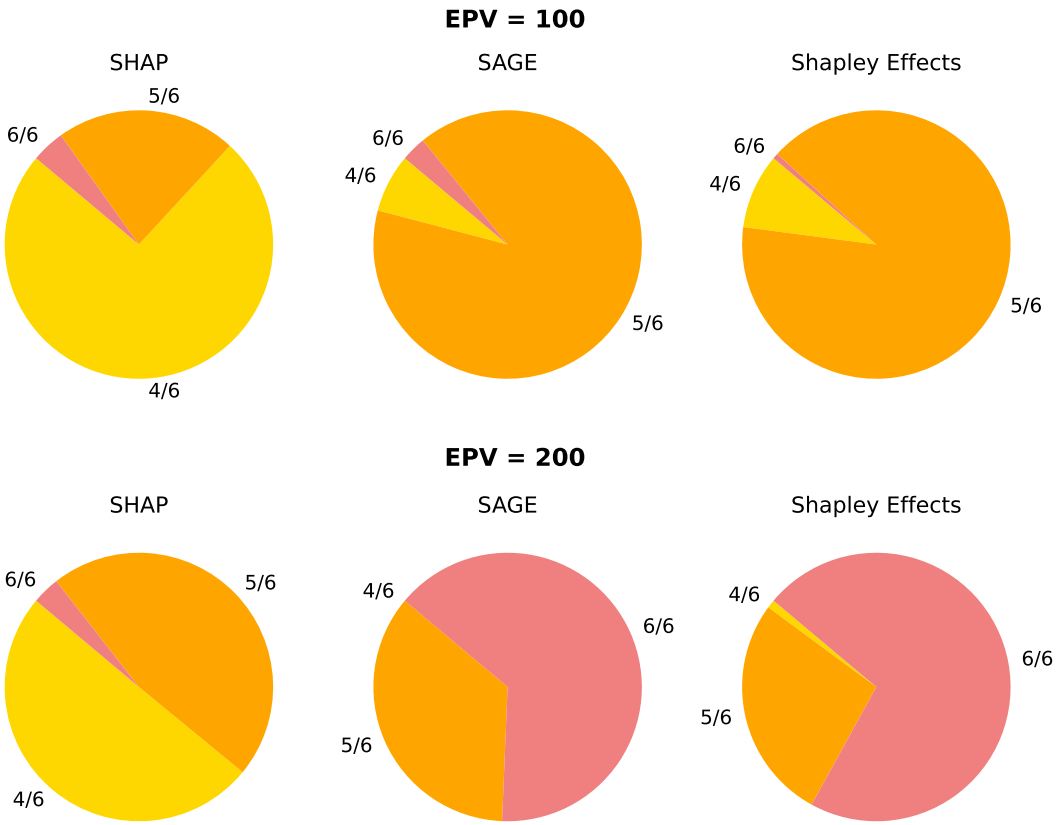

Figure 2: Frequency of correctly selected features for SAGE and Shapley effects, using XGBoost at $EPV = 100$ and $EPV = 200$. Mean absolute SHAP values for XGBoost have already been used in our prior work and serve as a reference [20]. $6/6$ implies that all informative features have been chosen, i.e., the correct feature subset has been recovered.

achieve excellent stabilities at 0.85 (95% CI [0.83, 0.87]) and 0.82 (95% CI [0.80, 0.84]). Greedy selection performs worst at 0.67 (95% CI [0.66, 0.68]). In terms of the discriminative performance, all strategies reach a mean test AUCs between 0.97 and 1.0 at $k = 2$, $k = 4$, and $k = 5$, and between 0.83 and 0.87 at $k = 3$.

For XGBoost, Shapley effects and SAGE also perform relatively well, albeit at much higher EPVs. At $EPV = 100$, the correct feature subset is almost never selected. In terms of the reliability, SAGE, Shapley effects, and SHAP perform similarly. With stabilities of only 0.74 (95% CI [0.73, 0.76]), 0.73 (95% CI [0.72, 0.75]), and 0.73 (95% CI [0.71, 0.75]), no strategy reaches excellent scores. At $EPV = 200$, though, Shapley effects and SAGE select the correct feature subset most of the time, whereas SHAP rarely recovers the correct features. With respect to the reliability, both Shapley effects and SAGE reach excellent stabilities at 0.83 (95% CI [0.81, 0.86]) and 0.81 (95% CI [0.79, 0.83]), whereas SHAP does not at 0.69 (95% CI [0.67, 0.71]). Regarding the discrimination, all methods achieve a mean test AUC of 1.0.

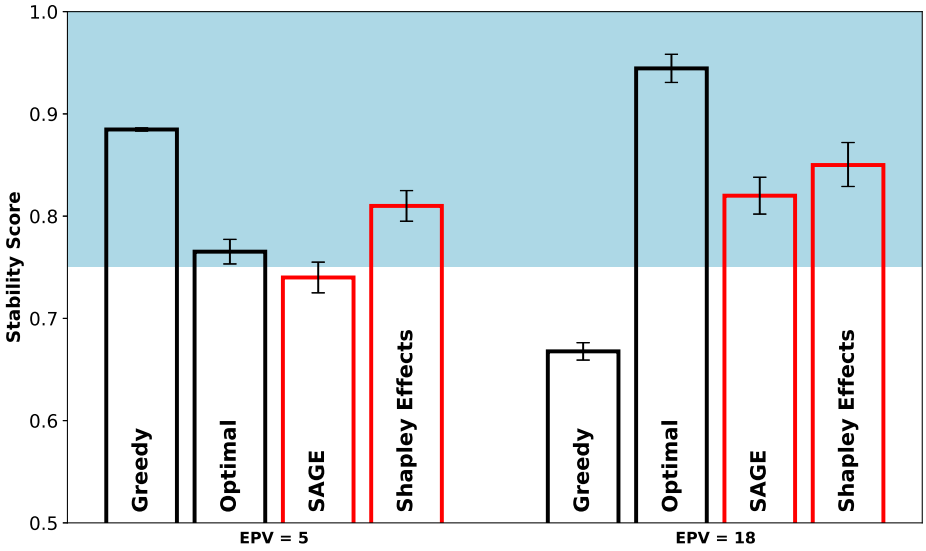

Figure 3: Stability or robustness to slight input perturbation for SAGE and Shapley effects, using logistic regression at $EPV = 5$ and $EPV = 18$. Greedy and optimal selection strategies for logistic regression have already been used in our prior work and serve as a reference [18]. The blue zone marks excellent stability scores [21].

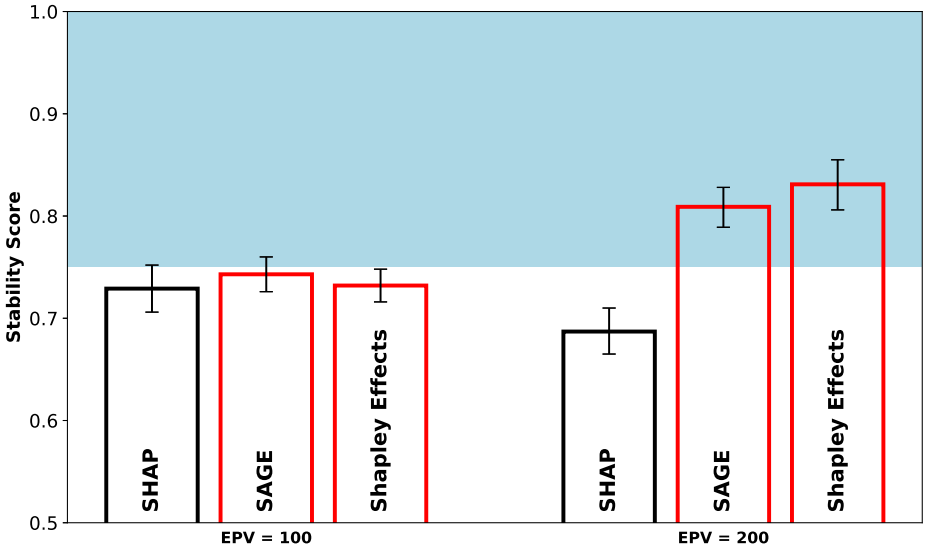

Figure 4: Stability or robustness to slight input perturbation for SAGE and Shapley effects, using XGBoost at $EPV = 100$ and $EPV = 200$. Mean absolute SHAP values for XGBoost have already been used in our prior work and serve as a reference [20]. The blue zone marks excellent stability scores [21].

4. Conclusion

Feature selection explanations are useful tools to understand which features are most influential for a given predictive model. In this work, we find that Shapley effects and SAGE values can offer

valid and reliable explanations given sufficiently large (algorithm-specific) EPVs. Increasing EPVs from 5 to 18 and 100 to 200 not only enhances validity but furthermore improves the stability from intermediate or good to excellent for these XAI methods. Nevertheless, this conclusion is based on only two common machine learning models for small- and medium-sized tabular data, L2-regularized logistic regression and XGBoost, on four synthetic datasets with two distinct data generating processes. Although these datasets provide clear ground truth explanations to evaluate XAI methods, further experiments are necessary to study the broader applicability of these methods for feature selection - for instance with respect to varying levels of correlations between features, missing values, or noise [18] and throughout various datasets and XAI benchmarks such as [22]. Additionally, it would be interesting to investigate feature selection explanations on even smaller sample sizes, where data-driven feature selection strategies may need to be complemented with prior knowledge in the form of causal graphs [23] or large language models [24]. We hope that future work will corroborate that XAI approaches are sufficiently valid and reliable to be applied across a variety of settings, can be trusted by practitioners, and can comply with regulatory standards that demand increasing levels of explainability and transparency for machine learning services.

Acknowledgments

This research was funded by the Bundesministerium für Bildung und Forschung (16DHBKI071).

References

[1] R. Adler, A. Bunte, S. Burton, J. Großmann, A. Jaschke, P. Kleen, J. M. Lorenz, J. Ma, K. Markert, H. Meeß, et al., Deutsche normungsroadmap künstliche intelligenz (2022).

[2] European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://artificialintelligenceact.eu/the-act/, 2021.

[3] I. Covert, S. Lundberg, S.-I. Lee, Explaining by removing: A unified framework for model explanation, Journal of Machine Learning Research 22 (2021) 1–90.

[4] A. B. Owen, Sobol'indices and shapley value, SIAM/ASA Journal on Uncertainty Quantification 2 (2014) 245–251.

[5] E. Song, B. L. Nelson, J. Staum, Shapley effects for global sensitivity analysis: Theory and computation, SIAM/ASA Journal on Uncertainty Quantification 4 (2016) 1060–1083.

[6] I. Covert, S. M. Lundberg, S.-I. Lee, Understanding global feature contributions with additive importance measures, Advances in Neural Information Processing Systems 33 (2020) 17212–17223.

[7] D. Fryer, I. Strümke, H. Nguyen, Shapley values for feature selection: The good, the bad, and the axioms, Ieee Access 9 (2021) 144352–144360.

[8] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[9] R. Knauer, E. Rodner, Squeezing lemons with hammers: An evaluation of automl and tabular

deep learning for data-scarce classification applications, arXiv preprint arXiv:2405.07662 (2024).
[10] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, Advances in neural information processing systems 35 (2022) 507–520.
[11] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, et al., Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion (2024) 102301.
[12] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, H. Lakkaraju, Rethinking stability for attribution-based explanations, arXiv preprint arXiv:2203.06877 (2022).
[13] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, Openxai: Towards a transparent evaluation of model explanations, Advances in Neural Information Processing Systems 35 (2022) 15784–15799.
[14] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods, arXiv preprint arXiv:1806.08049 (2018).
[15] H. Baniecki, P. Biecek, Manipulating shap via adversarial data perturbations (student abstract), Proceedings of the AAAI Conference on Artificial Intelligence (2022).
[16] F. E. Harrell, Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, Springer, 2015.
[17] G. Heinze, C. Wallisch, D. Dunkler, Variable selection–a review and recommendations for the practicing statistician, Biometrical journal 60 (2018) 431–449.
[18] R. Knauer, E. Rodner, Cost-sensitive best subset selection for logistic regression: A mixed-integer conic optimization perspective, in: German Conference on Artificial Intelligence (Künstliche Intelligenz), Springer, 2023, pp. 114–129.
[19] T. Van Der Ploeg, P. C. Austin, E. W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC medical research methodology 14 (2014) 1–13.
[20] R. Wallsberger, R. Knauer, S. Matzka, Explainable artificial intelligence in mechanical engineering: A synthetic dataset for comprehensive failure mode analysis, in: 2023 Fifth International Conference on Transdisciplinary AI (TransAI), IEEE, 2023, pp. 249–252.
[21] S. Nogueira, K. Sechidis, G. Brown, On the stability of feature selection algorithms, Journal of Machine Learning Research 18 (2018) 1–54.
[22] Y. Liu, S. Khandagale, C. White, W. Neiswanger, Synthetic benchmarks for scientific research in explainable machine learning, Advances in Neural Information Processing Systems (2021).
[23] T. Heskes, E. Sijben, I. G. Bucur, T. Claassen, Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models, Advances in neural information processing systems 33 (2020) 4778–4789.
[24] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, H. Lakkaraju, Are large language models post hoc explainers?, arXiv preprint arXiv:2310.05797 (2023).