CALIME: Causality-Aware Local Interpretable Model-Agnostic Explanations

Martina Cinquini[0000-0003-3101-3659] and Riccardo Guidotti[0000-0002-2827-7613]

University of Pisa, Italy
martina.cinquini@phd.unipi.it
riccardo.guidotti@unipi.it

Abstract. A significant drawback of eXplainable Artificial Intelligence (XAI) approaches is the assumption of feature independence. This paper focuses on integrating causal knowledge in XAI methods to increase trust and help users assess explanations' quality. We propose a novel extension to a widely used local and model-agnostic explainer that explicitly encodes causal relationships in the data generated around the input instance to explain. Extensive experiments show that our method achieves superior performance comparing the initial one for both the fidelity in mimicking the black-box and the stability of the explanations.

Keywords: Explainable Artificial Intelligence · Causal Discovery · Synthetic Data Generation · Interpretable Machine Learning

1 Introduction

In the past decade, eXplainable Artificial Intelligence (XAI) methods have been experiencing a wave of popularity due to their ability to provide human-understandable explanations that express the rationale of obscure models used by decision-making systems. According to the classification of XAI methodologies [7], we focus on post-hoc explainability. Given an AI decision system based on a black-box classifier and an instance to explain, post-hoc local explanation methods approximate the black-box behavior by learning an interpretable model in a synthetic neighborhood of the instance under analysis generated randomly. However, particular combinations of feature values might be unrealistic, leading to implausible synthetic instances [1]. This weakness emerges since, generally, these methods do not consider the local distribution features, the density of the class labels in the neighborhood [14], and, most importantly, the causal relationships among input features [10]. Such inability to disentangle correlation from causation can deliver sub-optimal or even erroneous explanations to decision-makers [20]. Causal Discovery (CD) methods can map all the causal pathways to a variable and infer how different variables are related. Even partial knowledge of the causal structure of observational data could improve understanding of which input features black-box models have used to make their predictions, allowing for a higher degree of interpretability and more robust explanations.

In this paper, we propose CALIME for Causality-Aware Local Interpretable Model-Agnostic Explanations. The method extends LIME [19] by accounting for underlying causal relationships present in the data on which the black-box operates. To attain this purpose, we replace the synthetic data generation performed by LIME through random sampling with GENCDA [4], a synthetic dataset generator for tabular data that explicitly allow for encoding causal dependencies. We highlight that our proposal of the causality-aware explanation method can be easily adapted to extend and improve other model-agnostic explainers like LORE [6], or SHAP [11]. However, in this work, we restrict our investigation and the integration of the knowledge of causal dependencies into the explanation extraction process of LIME due to the limitation of CD approaches that only work with continuous features and which are not widely adopted and developed to considering multiple attributes. Our experiments on various datasets show that CALIME significantly improves over LIME both for the stability of explanations and the fidelity in mimicking the black-box.

2 Related Works

LIME [19] is a model-agnostic method that returns local explanations as feature importance vectors. Further details are in Section 3. Despite the considerable number of LIME versions [18,21,2,22], no state-of-the-art variant explicitly allows causal relations to be encoded. However, using a causal system to extract explanations would rely on a more trustworthy and robust explanation process. XAI methods integrating causal knowledge are a challenging research area [15,12,9,16]. Even though some explainers account for causal relationships, to the best of our knowledge, no XAI techniques incorporate them during the explanation extraction. Some indirectly account for causality through a latent representation or adopt a known causal graph, typically as a post-hoc filtering step. Others consider causal relations between the input features and the outcome label but are not directly interested in the interactions among input features.

Hence, our proposal is innovative because it is the first in the panorama of post-hoc local explanation methods that directly consider causal relationships in the explanation extraction process. Besides, causal knowledge must not be provided a priori but is discovered by the explanation method itself.

3 Setting The Stage

We address here the *black-box outcome explanation problem* [7]. A classifier b is *black-box* when its internals are unknown to the observer, or they are known but uninterpretable by humans. Given b and an instance x classified by b, i.e., $b(x) = y$, the *black-box outcome explanation problem* aims at providing an explanation e belonging to a human-interpretable domain. According to the domain, in our work we focus on feature importance modeling the explanation as a vector $e = \{e_1, e_2, \dots, e_m\}$, in which the value $e_i \in e$ is the importance of the i^{th} feature for the decision made by $b(x)$. To understand the contribution of each feature, the sign and the magnitude of e_i are considered: if $e_i < 0$, the feature contributes negatively to the outcome y; otherwise, the feature contributes positively.

LIME. The main idea of LIME [19] is that the explanation may be derived locally from records generated randomly in the synthetic neighborhood Z of the instance x to be explained. It randomly draws samples and weights them w.r.t. a certain distance function π to capture the proximity with x. This gives to LIME a perturbed sample of instances Z to fed to the black-box b and obtain the classification probabilities $b_p(Z)$ with respect to the class $b(x) = y$. Such $b_p(Z)$ are combined with the weights W to train a linear regressor with Lasso regularization considering the top k most essential features. The coefficients of the linear regressor are returned as explanation e. A crucial weakness of LIME is the *randomness* of perturbations around the instance to be explained. To address this problem, we employ a data generation process that provides more realistic data respecting causal relationships.

GENCDA. In [4] is presented GENCDA, a synthetic data generator for tabular data that explicitly allows for encoding the causal structure among variables discovered by a Causal Discovery approach. In [4] is used NCDA, a boosted version of NCD [8]. Assuming there is no confounding, no selection bias, and no feedback among variables, NCD recovers the causal graph from the observational distribution by exploring a functional causal model in which effects are modeled as nonlinear functions of their causes and where the influence of the noise is restricted to be additive. GENCDA takes as input the *real* dataset X that has to be extended with *synthetic* data, the DAG G extracted from X by NCDA and a set of distributions. It returns a synthetic dataset compliant with the discovered causal relationships modeled in the DAG.

4 Causality-Aware LIME

In this section, we describe our method for Causality-Aware Local Interpretable Model-agnostic Explanations (CALIME). Before presenting CALIME, we illustrate through a working example why generating explanations without accounting for causality can be dangerous. Let consider a dataset X describing through *age*, *education level*, *income*, and *number of weekly working hours* the customers of a bank asking for loans. Let b be the AI system based on a black-box adopted by the bank and used to grant loans to customers $x \in X$. Now, let consider $x = \{(\text{age}, 24), (\text{edu\_level}, \text{high-school-5}), (\text{income}, 800), (\text{work\_hours}, 20)\}$ a customer that got the load denied. The bank wants to provide x with other reasons why the loan is denied and decides to apply LIME on x. A possible explanation for $b(x)$ could be $e = \{(\text{age}, 0.3), (\text{edu\_level}, 0.9), (\text{income}, 0.2), (\text{work\_hours}, 0)\}$. Thus, it seems for e that the low *education level* would be mainly responsible for the denied loan. By inspecting the neighborhood Z generated by LIME, we may discover many synthetic instances like $z = \{(\text{age}, 24), (\text{edu\_level}, \text{phd-8}), (\text{income}, 900), (\text{work\_hours}, 20)\}$ where a higher education level is present. If we had known the causal relationships among the customers in the training set of

Algorithm 1: CALIME(x, b, X, k, N)

Input : x - instance to explain, b - classifier, X - reference dataset, k - nbr of features, N - nbr of samples
Output: e - features importance

1 $Z \leftarrow \emptyset, W \leftarrow \emptyset$; `// init. empty synth data and weights`
2 $G \leftarrow \text{NCDA}(X)$; `// extract DAG modeling causal relationships`
3 $I \leftarrow \textit{fit\_distributions}(G, X)$; `// learn independent distributions`
4 $R \leftarrow \textit{fit\_regressors}(G, X)$; `// learn regressors`
5 **foreach** $i \in [1, \ldots, N]$ **do**
6     $z \leftarrow \text{GENCDA}\textit{\_sampling}(x, G, I, R)$; `// causal permutations`
7     $Z \leftarrow Z \cup \{z\}$; `// add synthetic instance`
8     $W \leftarrow W \cup \{exp(\frac{-\pi(x,z)^2}{\sigma^2})\}$; `// add weights`
9 $e \leftarrow \textit{solve\_Lasso}(Z, b_p(Z), W, k)$; `// get coefficients`
10 **return** e;

the black-box, we would have discovered that there is a relationship between *age* and *education level*, i.e., age → edu_level, and that synthetic instances like z are not plausible because they do not respect such relationship.

Therefore, the key idea of CALIME is to locally explain the predictions of any black-box model by also taking into account the underlying causal relationships in the data. The novelty of our proposal is actualized in *the generation of the neighborhood* around the instance to explain. Indeed, instead of random perturbation, CALIME employs GENCDA as a synthetic dataset generator for tabular data that discover nonlinear dependencies among the features and use them at generation time. Despite the apparent simple improvement of CALIME w.r.t., LIME, the ability of CALIME to encode explicitly causal relationships provides a considerable added value to explanations. In particular, respecting dependencies during the explanation process ensures that local explanations are based on plausible synthetic data and that the final explanations are more trustworthy and less subject to possible noise in the data.

The pseudo-code of CALIME is reported on Algorithm 1, the main differences with LIME are highlighted in blue[1]. First, CALIME runs on X the CD algorithm NCDA and extracts the DAG G that describes the causal structure of X (line 2). After that, for each *independent* variable j with respect to G, i.e., such that $pa(j) = \emptyset$, CALIME learns the best distribution that fits $X^{(j)}$ and produces a set of distribution generators I (line 3). For each *dependent* variable j concerning G, i.e., such that $pa(j) \neq \emptyset$, CALIME trains a regressor using as features $X^{pa(j)}$ and as target $X^{(j)}$, and produces a set of regressor generators R (line 4). After that, for each instance to generate, it locally runs GENCDA on x. Then, in line 6, CALIME takes as input the instance x, the DAG G, the set of distribution generators for independent variables I, and the set of regressor generators for dependent

[1] We highlight that the GENCDA algorithm does not explicitly appear in CALIME pseudo-code as it is decomposed among lines 2, 3, 4 and 6.

variables R. The GENCDA_*sampling* randomly selects the features $\{j_1, \dots, j_q\}$ to change among the independent ones[2]. For each independent feature to change j, the corresponding data generator I_j is used to generate a synthetic value z_j. After that, all the independent features have been synthetically generated, CALIME checks the causal relationships modeled in the DAG G. For all dependent variables j such that $pa(j)$ contains a variable that has been synthetically generated, the regressor generator R_j responsible for predicting the value of j is applied on $z^{pa(j)}$ to generate the updated value for feature j by respecting the causal relationship captured in G.

Let consider a dataset X describing through *age*, *education level*, *income*, and *number of weekly working hours*. Through NCDA, CALIME discovers a DAG G indicating the causal relation age $\rightarrow$ edu_level. Therefore, it learns the best distributions to model *age*, *income*, and *work_hours*. After that it learns a regressor $R_{\text{edu\_level}}$ on $\langle X^{(\text{age})}, X^{(\text{edu\_level})} \rangle$. Let consider the instance to explain $x = \{(\text{age}, 34), (\text{edu\_level}, 6.5),\ (\text{income}, 1000), (\text{work\_hours}, 35)\}$. When a synthetic instance z has to be generated, then *(i) education level* can not be changed if *age* is not changed, and *(ii)* when *age* is changed also *education level* must be changed according to $R_{\text{edu\_level}}$. For instance, if $z_{\text{age}} = 32$ then we can have $z_{\text{edu\_level}} = R_{\text{edu\_level}} = \text{phd-8}$. In this way, only synthetic customers with a higher *age* will be considered by the regressor when the *education level* is higher.

5 Experiments

We report here the experiments carried out to validate CALIME[3]. First, we illustrate the datasets used, the classifiers, and the experimental setup. Then, we present the evaluation measures adopted. Lastly, we demonstrate that our proposal outperforms LIME in terms of fidelity, plausibility, and stability.

5.1 Datasets and Classifiers

We experimented on datasets from UCI Repository[4] namely: `banknote`, `magic`, `wdbc`, `wine-red` and `statlog` which belong to diverse yet critical real-world applications. Table 1 (left) presents a summary of each dataset. Due to the nature of NCDA, all datasets have instances represented as continuous features. We plan as future research direction to extend CALIME with a different CD approach dealing also with categorical variables. We split each dataset into three partitions: X_b, is the set of records to train a black-box model; X_c, is the set of records reserved for discovering the causal relationships; X_t is the partition that contains the record to explain. We trained a Random Forest (`RF`) and a Neural Network (`NN`) as black box models[5]. For each black-box and dataset,

[2] The number of features to change q is selected uniformly at random in $[2, q]$.

[3] Python code and datasets available at: https://github.com/marti5ini/CALIME. Experiments were run on Ubuntu 20.04LTS, 252GB RAM, 3.30GHz x 36 IntelCore i9.

[4] https://archive.ics.uci.edu/ml/index.php. Accessed 6 Feb 2022.

[5] Black-boxes implementation: https://scikit-learn.org/. Accessed 3 Feb 2022.

Table 1. Dataset statistics and classifiers accuracy.

	n	m	l	X_b	X_c	X_t	RF	NN
`banknote`	1372	4	2	890	382	100	0.99	1.00
`statlog`	2310	19	7	1547	663	100	0.98	0.96
`magic`	19020	11	2	13244	5676	100	0.92	0.85
`wdbc`	569	30	2	328	141	100	0.95	0.92
`wine-red`	1159	11	6	358	203	154	0.82	0.70

we performed a random search for the best parameter setting[6]. Classification accuracy and partitioning sizes are shown in Table 1-(right).

5.2 Evaluation Measures

We evaluate the goodness of the explanations returned to three criteria, i.e., *fidelity*, *plausibility* and *stability* of the explanations.

Fidelity. We define the fidelity in terms of coefficient of determination $R_x^2 = 1 - (\sum_{i=1}^{N}(b(z_i) - r(z_i))^2)/(\sum_{i=1}^{N}(b(z_i) - \hat{y})^2)$ with $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} b(z_i)$ where $z_i \in Z$ is the synthetic neighborhood for a certain instance x, and r is the linear regressor with Lasso regularization trained on Z. R^2 ranges in $[-1, 1]$ and a value of 1 indicates that the regression predictions perfectly fit the data.

Plausibility. We evaluate the plausibility of the explanations in terms of the goodness of the synthetic datasets locally generated by LIME and CALIME. In [17] is presented a set of functionalities that facilitates the task of evaluating the quality of synthetic datasets. Our experiments exploit this framework to compare the synthetic data of the neighborhood Z with X_b.

Average Minimum Distance Metric. Given the local neighborhood Z generated around instance x, a synthetic instance $z_i \in Z$ is plausible if it is not too much different from the most similar instance in the reference dataset X. Given x, we calculate the plausibility in terms of Average Minimum Distance $AMD_x = \frac{1}{N}\sum_{i=1}^{N} d(z_i, \bar{x})$ with $\bar{x} = \arg\min_{x' \in X/\{x\}} d(z_i, x')$ where the lower the AMD, the more plausible are the instances in Z.

Outlier Detection Metrics. We also evaluate the plausibility of outliers in the synthetic neighborhood Z. The fewer outliers are in Z to X, the more reliable the explanation is. In particular, we estimate the number of outliers in Z by employing three outlier detection techniques[7]: Local Outlier Factor (LOF), Angle-Based Outlier Detection (ABOD), and Isolation Forest (IF) [3]. These three approaches return a value in $[0, N]$ indicating the number of outliers identified by the method. In the following, we report a single score named Average Outlier Scores (AOS) that combines the normalized scores of these indicators.

Statistical Metrics. To compare the real data X with Z we also used the Gaussian Mixture Log Likelihood, Inverted Kolmogorov-Smirnov D statistic, and

[6] Details of the parameters can be found in the repository.

[7] LOF and IF: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors. Accessed 15 Feb 2022. ABOD: https://pyod.readthedocs.io. Accessed 15 Feb 2022.

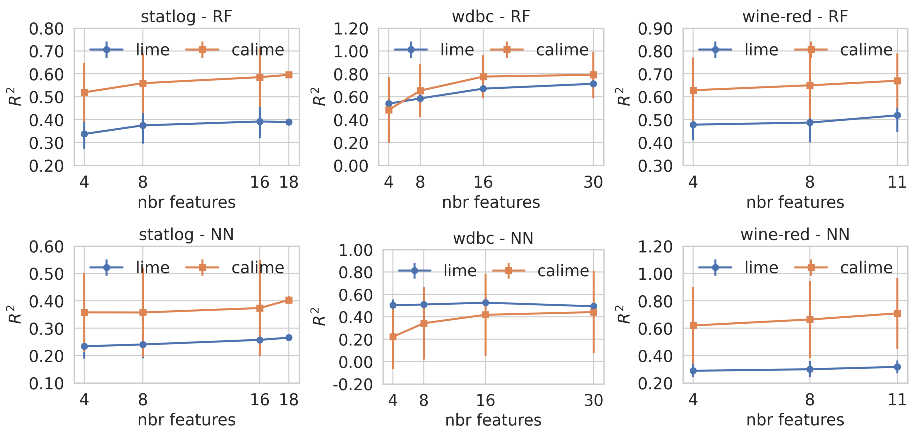

Fig. 1. Fidelity as $\boldsymbol{R^2}$ varying the number of features for `statlog`, `wdbc` and `wine-red`. Markers represents the mean values while vertical bars the standard deviations.

Continuous Kullback-Leibler Divergence [17]. In the results in the following, we report a single score named Average Statistical Metrics (ASM) that combines the normalized scores of the three indicators mentioned above.

Detection Metrics. The idea is to shuffle the real and synthetic data together, label them with flags indicating whether a specific record is real or synthetic, and cross-validate an ML classification model that tries to predict this flag [17]. We employ a Logistic detector and SVM as different "discriminators" and report the Average Detection Metric (ADM) of these approaches.

Stability. In line with [5], we asses the *stability* through the local Lipschitz estimation [13]: $LLE(x) = avg_{x_i \in \mathcal{N}_x^k}(\|e_i - e\|_2 / \|x_i - x\|_2)$ where x is the instance to explain and $\mathcal{N}_x^k \subset X$ is the k-Nearest Neighborhood of x with the k neighbors selected from X_t. The lower the LLE, the higher the stability.

5.3 Results

Due to space limitations, we report only significant results. Further analyses are available on the public repository. In Figure 1 we show the fidelity as R^2 of LIME and CALIME for `statlog`, `wdbc` and `wine-red` varying the number of features k. Results show that CALIME provides an improvement to LIME with the exception of `wdbc` for which the performance of CALIME are slightly worse than LIME. More in detail, for $k = 30$, LIME is more faithful for the NN classifier trained on `wdbc` since achieves .44 in contrast of .41 for CALIME. For the standard deviation, we notice that CALIME is less unstable than LIME. To summarize, CALIME provides explanations more trustworthy than those returned by LIME at the cost of a slightly larger variability in the fidelity of the local models.

In Figure 2 we illustrate the plausibility of LIME and CALIME through box plots aggregating the results obtained for the different number of features k for

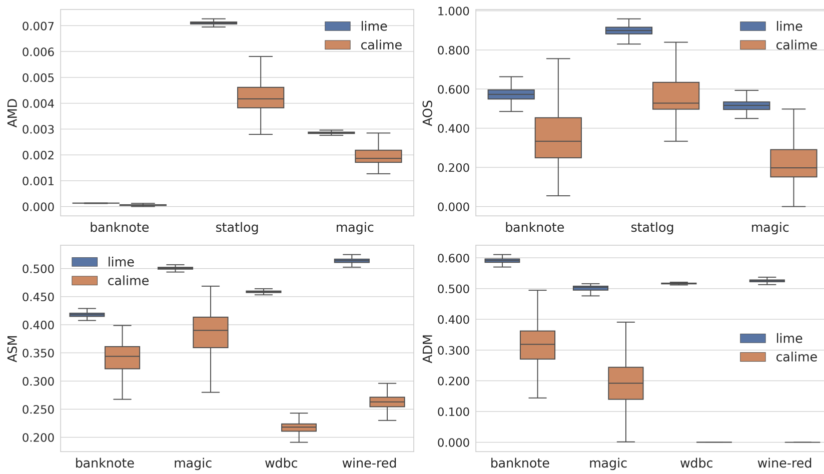

Fig. 2. Plausibility as AMD, AOD, ASM, and ADM in box-plots aggregating the results across the scores obtained with different number of features. Best view in color.

the selected datasets considering simultaneously results for the different black-boxes. We remind the reader that all the scores observed, i.e., AMD, AOD, ASM, and ADM, are "sort of" errors: lower values indicate higher plausibility scores. Analyzing the results, we observe that CALIME outperforms LIME in terms of plausibility for all the metrics analyzed. The improvement of CALIME can be attributed to learned relationships among the dataset variables that allow for generating data in the local neighborhood of x that more likely resembles the original ones. According to AMD and AOS scores, we observe the results focusing the analysis on `banknote`, `magic` and `statlog`. Overall CALIME generates a nearest and compact neighborhood. On the contrary, samples computed randomly by LIME are more distant from each other and have a low density around the instance to explain. For `banknote` we observe comparable results, while the more evident difference concerning AMD and AOS is highlighted for `statlog`. In terms of ASM and AMD, for which we report the insights w.r.t. `banknote`, `magic`, `wdbc` and `wine-red`, CALIME provides a more realistic and likely local synthetic neighborhood compared to LIME. The datasets that highlight the most remarkable improvement of LIME are `wdbc` and `wine-red`. Similar to what we notice for the fidelity, the highest plausibility of CALIME is paid with a more variegated neighborhood, shown by the largest box plots (e.g., for `magic`).

In Figure 3 we illustrate the stability of LIME and CALIME as LLE (the lower, the better) for `banknote`, `wine-red` and `wdbc` varying the number of features k. The colored areas highlight the minimum and maximum values of LLE obtained by replacing the min and max operators in the previous formula. We notice that LIME is more stable than CALIME or comparable when the number of features perturbed by the neighborhood generation procedures is small, i.e., $k \leq 4$. On the other hand, CALIME is more resistant to noise than LIME when $k > 4$.

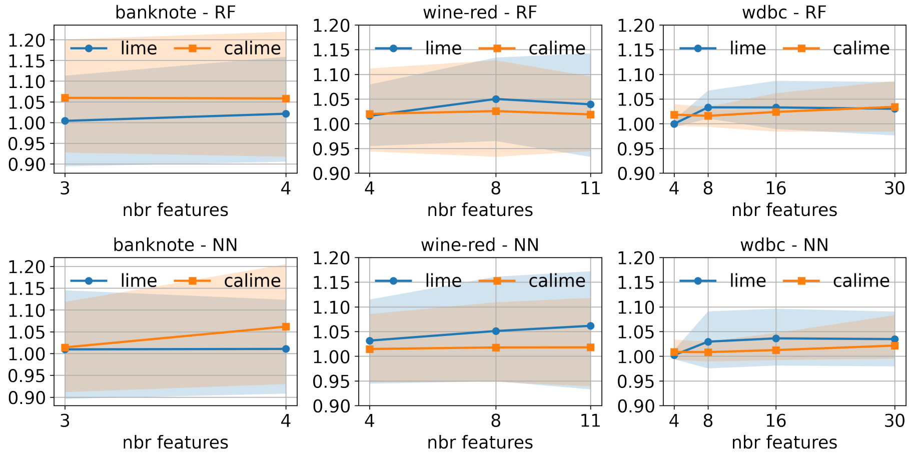

Fig. 3. Stability as $\boldsymbol{LLE}$ varying the number of features. Markers represents the mean values while the contingency area highlights the minimum and maximum values.

In summary, CALIME empirically exhibits better performance than LIME on the datasets and for the black-boxes analyzed. A drawback of CALIME, at least concerning the current implementation, is that it is considerably slower due to the overhead brought by the time required *(i)* for extracting the DAG and *(ii)* for learning the distributions of the independent variables and the regressors to approximate the dependent variables. LIME explanations can be returned in less than a second, at least for the datasets and classifiers analyzed in this study, while CALIME requires one to two more orders of magnitudes.

6 Conclusion

We have presented the first proposal in the research area of post-hoc local model-agnostic explanation methods that *discovers* and *incorporates* causal relationships in the explanation extraction process. Due to the mapping of explainability with causality, the exploitation of CALIME strengthens user trust in the AI system. A limitation of our proposal is that adopting GENCDA that in turn is based on NCDA, CALIME can only work on datasets composed of continuous features. To overcome this drawback, we need to rely on CD approaches to simultaneously account for heterogeneous continuous and categorical datasets. Finally, to completely cover LIME applicability, we would like to study to which extent it is possible to employ causality awareness on data types different from tabular data such as images and time series.

Acknowledgments. This work has been partially supported by the EU H2020 program under the funding schemes: G.A. 871042 *SoBigData++*, G.A. 952026 *HumanE-AI Net*, ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making*, G.A. 952215 *TAILOR*, G.A. CHIST-ERA-19-XAI-010 *SAI*.

References

1. A. Artelt, A. Vaquet, V. Velioglu, R. Hinder, F. Brinkrolf, and J. Schilling. Evaluating robustness of counterfactual explanations. *CoRR*, abs/2103.02354, 2021.
2. S. Bramhall, H. Horn, M. Tieu, and N. Lohia. Qlime-a quadratic local interpretable model-agnostic explanation approach. *SMU Data Science Review*, 3(1):4, 2020.
3. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.
4. M. Cinquini, F. Giannotti, and R. Guidotti. Boosting synthetic data generation with effective nonlinear causal discovery. In *CogMI*, pages 54–63. IEEE, 2021.
5. R. Guidotti et al. Black box explanation by learning image exemplars in the latent feature space. In *ECML/PKDD*, volume 1 of *LNCS*, pages 189–205. Springer, 2019.
6. R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, and S. Ruggieri. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.*, 34(6):14–23, 2019.
7. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.
8. P. O. Hoyer et al. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696. Curran Associates, Inc., 2008.
9. A. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *FAccT*, pages 353–362. ACM, 2021.
10. L. Longo et al. Xai: Concepts, applications, research challenges and visions. In *CD-MAKE*, volume 1 of *LNCS*, pages 1–16. Springer, 2020.
11. S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.
12. D. Mahajan, C. Tan, and A. Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *CoRR*, abs/1912.03277, 2019.
13. D. A. Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.
14. M. Moradi and M. Samwald. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, 165:113941, 2021.
15. R. Moraffah et al. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.
16. A. Parafita and J. Vitrià. Explaining visual models by causal attribution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4167–4175. IEEE, 2019.
17. N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *DSAA*, pages 399–410. IEEE, 2016.
18. T. Peltola. Local interpretable model-agnostic explanations of bayesian predictive models via kullback-leibler projections. *CoRR*, abs/1810.02678, 2018.
19. M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *KDD*, pages 1135–1144. ACM, 2016.
20. J. G. Richens, C. M. Lee, and S. Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.
21. S. M. Shankaranarayana et al. ALIME: autoencoder based approach for local interpretability. In *IDEAL*, volume 11871 of *LNCS*, pages 454–463. Springer, 2019.
22. X. Zhao, W. Huang, X. Huang, V. Robu, and D. Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. In *Uncertainty in Artificial Intelligence*, pages 887–896. PMLR, 2021.