Causality Learning: A New Perspective for Interpretable Machine Learning

Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu and Xianzhi Wang

Abstract—Recent years have witnessed the rapid growth of machine learning in a wide range of fields such as image recognition, text classification, credit scoring prediction, recommendation system, etc. In spite of their great performance in different sectors, researchers still concern about the mechanism under any machine learning (ML) techniques that are inherently black-box and becoming more complex to achieve higher accuracy. Therefore, interpreting machine learning model is currently a mainstream topic in the research community. However, the traditional interpretable machine learning focuses on the association instead of the causality. This paper provides an overview of causal analysis with the fundamental background and key concepts, and then summarizes most recent causal approaches for interpretable machine learning. The evaluation techniques for assessing method quality, and open problems in causal interpretability are also discussed in this paper.

Index Terms—Interpretable Machine Learning, Causal Inference, Counterfactual Explanation, Causal Feature, Causal Interpretability

I. INTRODUCTION

IN the past decades, machine learning has achieved the impressive performance in diverse tasks, and is increasingly applied in science, society and business. However, most of state-of-the-art models remained incomprehensible for both researchers, users and engineers, causing difficulties when deploying in real world. Specifically, there are several high-stake decision-making domains such as self-driving cars, crime prediction or personalized medicine in which the lack of transparency in machine learning prevents themselves from being adopted. Take for instance, in the healthcare sector where each decision can affect the people's survival, physicians are frequently concerned about the safety and trust of any deployed models. They do not likely trust the model's prediction if they can not understanding the rationales behind it. Consequently, interpretability in machine learning plays a significantly important role in generating trust-worthy models. This furthermore allows researchers, data scientists and engineers to ensure the models following the human understanding, ethnic codes, fairness and security. We as human have an insatiable curious nature; thus, our goal is not only to understand models' mechanism but also to generate and extract new knowledge of the world.

In view of the time of explanation generation shown in Figure 1, interpretable machine learning can be divided into two branches: ad-hoc and post-hoc methods. The evolutionary

School of Computer Science, University of Technology Sydney, Australia [Guandong.Xu,Qian.Li,Shaowu.Liu,Xianzhi.Wang]uts.edu.au, TriDung.Duong@student.uts.edu.au

history of noticeable traditional interpretable machine learning techniques is briefly described in the Figure 1. The ad-hoc type focuses on building the model architecture, algorithms or mechanisms that are self-explainable and transparent. Intrinsically interpretable models are the central research in the early years of artificial intelligence with the dominance of symbolism methods, followed by more advanced approaches such as decision sets [1], generalized linear regression, generalized additive model [2], [3], [4], Bayesian probabilistic model [5], [6], rule-based model [7], [8], attention mechanism [9], fuzzy inference systems [10], [11], [12], TabNet [9], etc. With the rapid growth of deep learning in recent decades, machine learning model is gradually evolved into complicated and incomprehensible form, which leads to the increasing attention on post-hoc interpretations. Several prominent approaches in this category include Local surrogate models (LIME [13], SHAP [14], LORE [15], etc), influence functions [16] and feature importance estimation [17], [18] have been introduced.

However, traditional interpretable machine learning focuses on the association instead of the causality. With the emergence of causal inference, an increasing number of causality-oriented methods have been proposed in interpretable machine learning. In comparison with traditional methods, causal approaches can be utilized to identify causes and effects of models architecture or conduct the reasoning over its decisions and behaviors. This article examines the overview of interpretable machine learning, presents the causal analysis in machine learning interpretability and finally discusses the future research directions. More specifically, we first present the background of causal analysis with key concepts, models and evaluation metrics. We then provide an overview of state-of-the-art works on causal interpretability. We also illustrate the potential evaluation metrics used in interpretable machine learning.

II. CAUSALITY ANALYSIS

Causality analysis can exploit the causality mechanisms underlying the data-generating process, which is more advanced than the predictive or descriptive capability in machine learning techniques. Causal inference and causal discovery are two main research topics for causality analysis. The goal of causal inference is to estimate the causal effect of treatment (i.e., a decision made or action taken) on the outcome (i.e., the result of treatment). Causal discovery examines whether a set of causal relationships exists among the variables. This paper would primarily focus on causal effect, which is more correlated to machine learning interpretability.
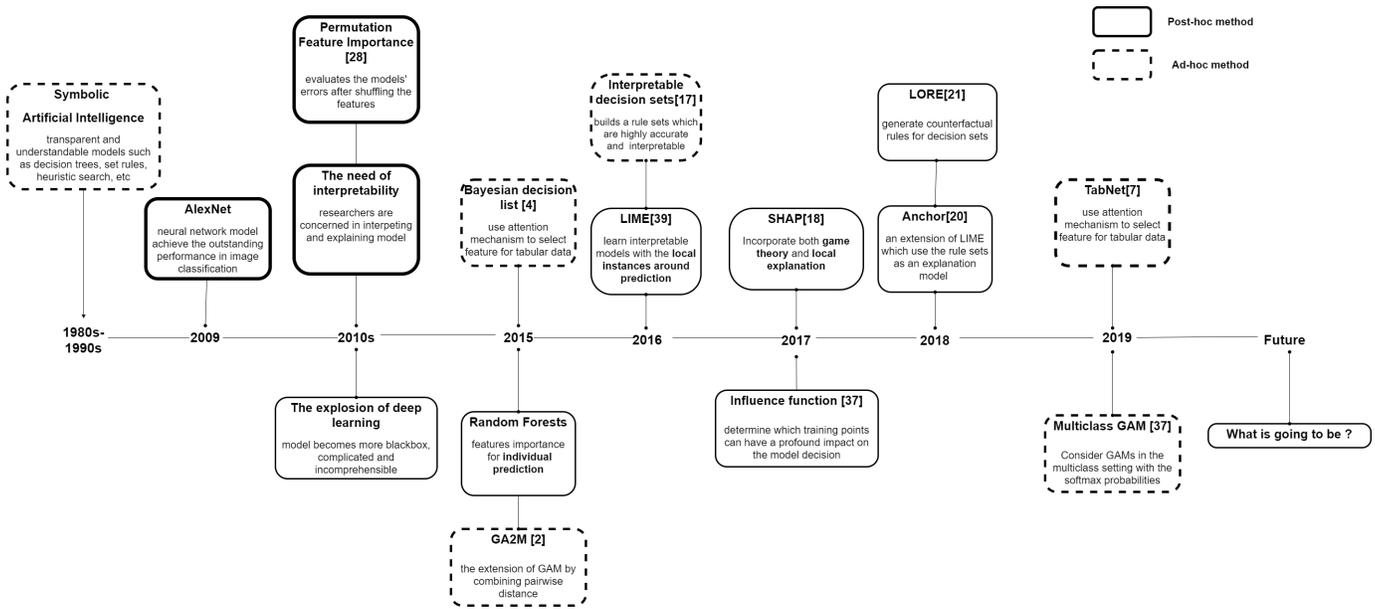
Fig. 1.   The evolution of interpretable machine learning

A. Causal Inference

Causal inference has been widely applied in econometric, social science and medicine fields for evaluating the policy's effect or the drugs' side effect. Effect estimation is tied to the outcome caused by the treatment applied to an instance. An instance is the atomic research object, which can be a physical object or an individual person. Treatment and outcome are terms that denote a decision made or action taken and its result, respectively. We first introduce the essential concepts for learning treatment effect followed by the causal models.

- Covariates X refers to the background variables or features of the instance.
- Treatment T refers to the action (manipulation or intervention) that applies to a instance.
- Outcome Y is the result of the treatment applied on a instance.
- Confounder Z is a variable which causally affects both treatment and outcome.

To better understand causal inference, we give the following example combined with the notations defined above. To prove the efficiency of the medication on the disease, the scientist needs to assess its positive effect into the patients' recovery rate. Figure 2 depicts the corresponding causal relationships among the essential variables. The treatment T is whether the drugs are applied or not, and the observed features X are the patients' condition such as the level of insulin and cholesterol, heart rate, etc. Outcome Y is the recovery rate and age is the confounder Z. This is simply because age firstly determined the need of applying medication into patients, since the young people may not necessarily take the medicine. Age also affects to the recovery rate: the youth has a higher probability to recover than the elderly.
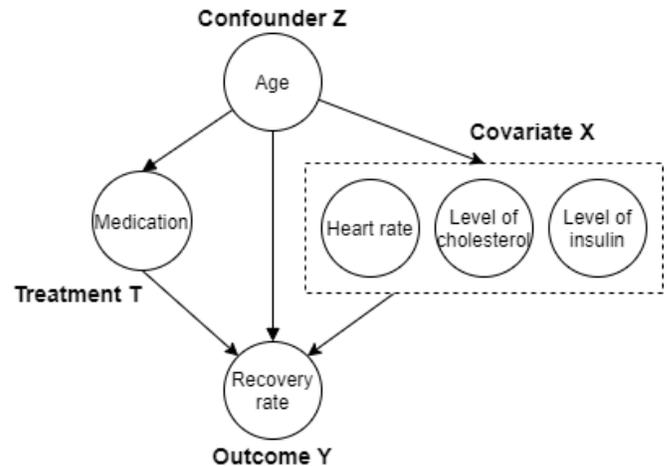

Fig. 2.   The causal graph for recovery rate problem

B. Causal Models

We now introduce the two most important formal frameworks used for causal inference, namely the structural causal models and the potential outcome framework.

Structural causal model[19] consists of two main components: the causal graph and structural equations. Causal graph is the probabilistic graphical model which is used to represent the assumption about prior knowledge and data generating process. A causal graph is defined as $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Structural equation is a set of equations Eq. (1) which are used to represent the causal effect illustrated by the edge in the causal graph.

$$X = f_X(E_X),$$
$$T = f_T(X, E_T). \qquad (1)$$
$$Y = f_Y(X, D, E_Y)$$

where E_X, E_T, E_Y are exogenous variables, which are in-

dependent from other models' variable, and are determined outside the model.

Potential outcome framework is proposed by Neyman and Rubin [20]. Considering binary treatments for a set of units, there are two possible outcomes for each unit. The unit will be assigned to the control treatment if $T = 0$, or to the treated treatment if $T = 1$. As a result, we denote two potential outcomes Y_0 and Y_1 as the results caused by $T = 0$ and $T = 1$, respectively. Importantly, only one potential outcome is observed corresponds to the assigned treatment T, and we call this as the observed (factual) outcome Y. The unobserved potential outcome refers to the counterfactual outcome. Given the treatment T_i, the relationship between the observed outcome Y and two potential outcomes are

$$Y_i = T_i Y_1 + (1 - T_i) Y_0 \tag{2}$$

C. Treatment Effect Metric

With the key concepts and causal models, the treatment effect can be measured at the population, treated group, subgroup, and individual level. For simplicity, we discuss the treatment effect under the binary treatment, and it can be easily extended to multiple treatments by considering multiple potential outcomes.

The individual treatment effect (ITE) is defined as the change of Y_0 and Y_1, while keeping the covariates X unchanged (i.e., condition on those covariates). For an instance i with covariates X_i, its corresponding ITE is

$$ITE(\boldsymbol{X}_i) = E[Y_1|X_i] - E[Y_0|X_i] \tag{3}$$

As only one potential outcome is observed, it is nearly impossible to estimate the effect at the individual level. A more feasible way is to measure treatment effect at the average level.

The average treatment effect (ATE) measures the treatment effect at the whole population level as

$$ATE = E[Y_1 - Y_0] \tag{4}$$

The average treatment effect (ATT) is for the group of instances with the treatment equal to 1, i.e., the treated group.

$$ATT = E[Y_1 - Y_0|T = 1] \tag{5}$$

Conditional average treatment effect (CATE) known as heterogeneous treatment effect is defined on the subgroup with the particular covariate $X = x$.

$$CATE = E[Y_1 - Y_0|X = x] \tag{6}$$

D. Tools for Causal Analysis

Several libraries or tools are available for causal inference. Examples including *Double Machine Learning* [21], *Meta-learners* [22], *Orthogonal Learning* [23], [24] have been supported by EconML, CausalML, DoWhy and CausalNex, whereas causal discovery methods including graph inference and pairwise inference are provided in Causal Discovery Toolbox. Meanwhile, TIGRAMITE is a novel framework for causal discovery in time series. We summarize the existing toolboxes in Table I.

III. INTERPRETABLE MACHINE LEARNING WITH CAUSALITY

Pearl [25] argues that causal reasoning is indispensable for machine learning to reach human-level artificial intelligence, since it is the basic mechanism for humans to be aware of the world. As a result, causal methodology is gradually becoming a vital component in explainable and interpretable machine learning. However, most of current interpretability techniques pay attention to solving the correlation statistic rather than the causation. Therefore, the causal approaches should be emphasized to achieve a higher degree of interpretability.

A. Model-Agnostic Causality for Deep Neural Neworks

The traditional way to analyze Deep Neural Network is to build several models with different architectures and make a comparison between their performances. The problem is that re-training DNNs is computationally expensive, and infeasible when it comes to the complicated architecture. Inspired by causal model, several methods have been proposed to interpret neural network model.

Chattopadhyay et al. [26] define $\text{ACE}^y_{do(x_i=\alpha)}$ as the causal attribution of neuron x_i to the output neuron y_i, and $\mathbb{E}[y|do(x_i = \alpha)]$ as the interventional expectation Eq. (7). The polynomial function is selected to estimate this value.

$$\mathbb{E}[y|do(x_i = \alpha)] = \int y p(y|do(x_i = \alpha)) dy \tag{7}$$

Narendra et al. [27] propose to construct a modified structural causal model as an abstraction of a DNN to make an reasoning over its elements. Thereafter, they rank each component based on their contribution to the final prediction for evaluation.

Based on TCAV [28] which generates a high-level concept-based explanation such as gender, race, background, others, the study in [29] evaluates the *causal concept effect* on a neural network prediction. They overcome the problem of do-operator by using Variational AutoEncoder (VAE).

Regarding Generative Adversarial Networks (GANs) interpretability, Bau et al. [30] proposes an approach for visualization and understanding at unit-, object-, and scene- level by estimating the causal effect of the models' interpretable components. There are two main steps in their approach: dissection and intervention. In the dissection step, the classes with the explicit representation are firstly identified. Thereafter, they make an intervention by forcing the units to be appeared and disappeared, and calculate its causal effect. Meanwhile, the authors [31] propose a causal framework to explore the intervention effect for proving that the components in images generated by GAN can be modified independently.

In terms of reinforcement learning, action influence model [32] is introduced for explaining the behavior of RL agents. They construct a modified structural causal model, learn the causal equation as the regression model during training the agent, and finally generate the contrastive explanation to answer the counterfactual question "Why does the agent choose action A instead of action B?".

B. Post-hoc Interpretability

Model-Agnostic explanations are particular challenging when the models' parameters have more complex relationships. To further aid the intepretability, the practitioners propose a variety of post-hoc interpretability methods to exploit what a trained model has learned, without changing the underlying model. Most widely useful post-hoc interpretation methods fall into two main categories: causal feature learning and counterfactual explanations, respectively.

1) Casual Feature Learning: Recent work on feature learning derives the subset of features that have causal contributions to the models' prediction. Early causal feature learning is to find the Markov Blanket (MB) containing a set of features which makes the target (T) independent from other features given MB(T). In the study [33], the authors firstly use the HITON algorithm [34] to derive the Markov Blanket, and thereafter deploy Max-Min Hill-Climbing (MMHC) algorithm to identify the causes and effects of the target variable. Given the number of transfer learning tasks D, Peters et al. [35] assume that there exists a subset of features X_{S*} such that the conditional distribution $Y_k|X_{S*}$ is the same for different tasks k, and other settings Eq. (8). They propose an algorithm called subset search which samples the subset features, and then adopt the Levene test to assess the assumption.

$$Y_k|X_{S*}^k \approx Y_k'|X_{S*}^{k'} \quad \forall k, k' \in 1, ..., D \tag{8}$$

CXPlain [18] is the causal framework that can explain more complex machine learning models by estimating the feature importance. Granger-causal objective is introduced to quantify how much the exclusion of a single feature reduces model performance. Particularly, CXPlain trains a separate explanation model to any predictor f by optimizing a Granger-causal objective. CXPlain can also estimate the uncertainty of features importance by calculating confidence interval (CI).

2) Counterfactual Explanation: Counterfactual explanation is the example-based model-agnostic method which generates new instances that would change the models' prediction. The prominent example [36] in this research is that a person x with an annual income a and a current balance b has been rejected a loan by the financial institution, so how can they change their income and balance to a' and b' to receive the loan. Given the set of points P, to generate the set of counterfactual samples F, the objective function of counterfactual explanation [37] is to optimize the following function:

$$\arg \min_x \max_\lambda (\lambda \cdot (\hat{f}(x') - y')^2 + d(x, x'))$$
$$d(x_i, x') = \sum_{k \in F} \frac{|x_k - x_k'|}{MAD_k} \tag{9}$$
$$MAD_k = \underset{(j \in P)}{\text{median}}(|X_{j,k} - \underset{(l \in P)}{\text{median}}(X_{l,k})|)$$

where x is an original instance, x' is the counterfactual instance which close to x, y' is the target class label for x', λ is the regularized parameter, $d(x, x')$ denotes the distance between the original instance and the counterfactual samples, MAD_k is the median absolute deviation for feature k.

Grath et al. [36] extend $d(x, x')$ in Eq. (9) by adding a weight vector Θ. The vector Θ is used to evaluate models' feature importance, and can be obtained by many algorithms such as K-Nearest Neighbors or global feature evaluation. Dhurandhar et al. [38] combine the loss function generated from Convolutional AutoEncoder, while Arnaud [39] uses the prototypes function to ensure that the generated perturbation falls into the same distribution with the original data as well as increasing the computational speed without tuning too many parameters. Additionally, the counterfactual samples should be as diverse as possible; the study [40] proposes to use determinant of kernel matrix to illustrate this property.

To empower the capability of counterfactual explanations, constraints are considered in optimization problem of counterfactual explanation. Take for example, a person cannot decrease his age, or change his race and skin color. Recent work [41], [42] adopt Mixed Integer Programming (MIP) formulation to deal with categorical, numeric and mixed data type. Meanwhile, Artelt et al. [43] propose convex density constraints to generate counterfactual located in a region of the data space. Specifically, the density constraint $\hat{p}_y \geq \delta$ denoted by a kernel density estimator or a Gaussian mixture model is added into the distance function $d(x, x')$.

CERTIFAI [44] proposed by Sharma et al. as a novel and flexible approach which can be used in any type of data. CERTIFAI uses the customized genetic algorithm to choose individuals that have the best fitness scores defined as follows.

$$fitness = \frac{1}{d(x, x')}$$
$$d(x, x') = \begin{cases} \frac{n_{x'on}}{n} l_1(\mathbf{x}, \mathbf{x}') + \frac{n_{cat}}{n} simp(\mathbf{x}, \mathbf{x}') & \text{tabular data} \\ \frac{1}{SSIM(x,x')} & \text{image data} \end{cases} \tag{10}$$

For tabular data, CERTIFAI chooses l_1 norm for continuous features and a simple matching distance for categorical features (simp). For image data, Structural Similarity Index Measure (SSIM) [45] measures the similarity of what humans consider. n_{con} and n_{cat} are the number of continuous features and categorical features, respectively.

Instead of identifying the minimum changes leading to the desired outcome, a new line of counterfactual explanations provides feasible paths to transform a selected instance into one that meets a certain goal. FACE [46] proposed by Poyiadzi et al. constructs a graph over the data points with the weights illustrating the feasible degree to transit between two vertices. FACE thereafter can be solved by the *Disjstra* algorithm to find the shortest path from the original instance to the counterfactual one.

C. Visualization of Causal Effect

Visualization-based method is another commonplace approach for quick understanding what the models have learned. Partial dependence plot (PDP) [47] depicts the marginal effect of features into the predicted outcomes. The partial dependence function is defined as:

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)p(x_C)dx_C \tag{11}$$

Zhao et al.[48] use Partial dependence plot (PDP) an its extension called Individual Conditional Expectation (ICE) to extract the causal information from machine learning model. These visualization tools allow to measure the predictions' change after making an intervention, which can help to discover the features' causal relationship.

IV. Evaluation

Evaluation in causal interpretability is an extremely difficult task currently, since there are nearly no groundtruth data to evaluate the methods' performance. Evaluation for traditional interpretable machine learning evaluation can be classified into three categories [49]: application-based, human-based and function-based. We apply the same category and focus on evaluations that can be used in causal interpretability.

A. Application-based

In real-world scenario where the machine learning model is deployed to assist experts, application-based evaluation illustrates how well the models provide explanations to human experts for improving their performance in specific tasks. Take for example, a randomized experiment [50] is conducted among a group of learner to solve the problems. They then rate the explanation generated by the machine learning models. With the assistance of models, the performance of people in different tasks is proved to be improved.

B. Human-based

Human-based evaluation methods refer to evaluate the performance of interpretable models with the assistance of human. Madumal et al. [32] generate explanation for the reinforcement learning. They implement an RL agent, and conduct an experiment running on StarCraft II, a strategic game, with 120 participants. *Explanation Satisfactory Scale* [51] is defined as the degree of human understanding of the AI system to measure the quality of generated explanations.

C. Function-based

Functional-based evaluation methods can be carried out without the assistance of human to evaluate the performance of the explanation model. There are some evaluation procedures for different techniques in Section IV:

1) Causal Interpretability for DNN: The lack of ground truth for feature effect makes it challenging to evaluate the performance of causal effect estimation. Chattopadhyay et al. [26] compare the salient map [52] generated by causal attribution method with Integrated Gradient [53]. Harradon etc al. [54] identify the components having the significant causal effects into the individual prediction. Specifically, they conduct the experiments in three different architectures VGG 19 in Birds200, VGG 16 and 6-layer cov network applied in Inria dataset. Thereafter, they make a query for an individual input, and then visualize top k variables according to their causal effect.

2) Counterfactual Explanations: A previous research [40] suggests that there are three main metrics to evaluate the counterfactual explanation: *proximity*, *diversity* and *sparsity*. The *proximity* is to reflect the similarity between the CF examples and the original one which was calculated as the mean proximity all over the examples. Meanwhile, the *diversity* measures the mean of the distances between the pairs of samples, ensuring that the generated instances should be as diverse as possible. Finally, the *sparsity* is the average number of changes converting CF examples to the original one.

$$proximity = \frac{-1}{k}\sum_{i=1}^{k} dist(x_{cf_i}, x)$$

$$diversity = \frac{1}{C_2^k}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} dist(x_{cf_i}, x_{cf_j}) \tag{12}$$

$$sparsity = 1 - \frac{1}{k \cdot d}\sum_{i=1}^{k}\sum_{l=1}^{d} 1[x_{cf_i}^l \neq x_i^l]$$

with x_{cf} and x are the counterfactual samples and original instance, respectively, $dist(x_{cf_i}, x_{cf_j})$ illustrates the distance between two generated counterfactual instances, d is the number of input features, k is the number of counterfactual samples to be generated.

V. Open questions and discussions

The need of explaining and interpreting models becomes highly critical along with the growing popularity of deep learning and automated machine learning. Although, there are currently several studies in this field, several open problems still remain unresolved.

1) Counterfactual explanation in classification tasks. There are a plethora of constraints, especially features' causal relationship, should be taken into consideration when adopting counterfactual explanation. Take for example, the counterfactual explanation cannot recommend the users to change sensitive and discriminative features such as race and gender in order to be accepted by the system. Therefore, its reasonability and feasibility should be discovered and investigated more strictly.

2) Counterfactual explanation in recommendation system and time series data. Although recommendation system gains the immense popularity these days, there are not many studies working on counterfactual explanation for such system. How we can make an intervention into human actions to enable the system to change their recommended items still remains an open question. Meanwhile, regarding time series data, it is also interesting to discover that what the model would change its prediction if we change something in the past.

3) Causal reasoning in knowledge graph. Knowledge graph is recently utilized as an effective tool in several tasks such as recommendation system, knowledge extraction, classification, etc. Instead of embedding the knowledge graph as the latent features, Xian et al. [68] state that the true intelligent recommendation systems have to own the ability to recommend their items based on their causal reasoning.

4) Explanation understandable by non-experts. A number of recent methods frequently provide the explanations to

TABLE I
TOOLBOX FOR CAUSAL ANALYSIS

Library	Feature	Algorithms	License
DoWhy [55]	Individual treatment effect estimation	Prospensity score matching [56]	
Stratification [57]	Microsoft		
EconML [58]	Individual treatment effect estimation		
Interpreter of the causal model	Double Machine Learning [21]		
Orthogonal Random Forests [23], [24]			
Meta-Learners [22]			
Deep Instrumental Variables	Microsoft		
Causal ML [59]	Individual treatment effect estimation	Meta-learners	
Uplift modeling [60], [61]	Uber		
Causal discovery toolbox [62]	Causal relationship discovery	Graph Inference	
Pairwise inference	ElementAI		
CausalNex	Learn causal structures		
Estimate the effects of potential interventions using data.	Using Bayesian Networks for Causal Inference	QuantumBlack Labs	
TIGRAMITE	Causal discovery for time series datasets	PCMCI [63], Generally [64], CMIknn [65], Mediation class [66], [67]	GNU General Public

experts and researchers rather than the end-users. Therefore, another challenge is to generate explanation under the form such as rules, natural language, images, etc which can allow nonprofessional people to catch up with machine learning model behaviors.

VI. CONCLUSION

Interpretable machine learning is expected to become a mainstream topic in the foreseeable future. This paper provides the desiderata and brief overview of causal inference, followed by the causality based interpretable machine learning. We present two main causal approaches for interpretable machine learning including feature importance estimation, causal effects of model components, and counterfactual explanation. Finally, we has discussed several potentially unresolved problems in this field which open opportunities for researchers to work in.

In machine learning, the more data the better. However, in causal inference, the more data alone is not yet enough. Having more data only helps to get more precise estimates, but it cannot make sure these estimates are correct and unbiased. Machine learning methods enhance the development of causal inference, meanwhile, causal inference also helps machine learning methods. The simple pursuit of predictive accuracy is insufficient for modern machine learning research, and correctness and interpretability are also the targets of machine learning methods. Causal inference is starting to help to improve machine learning, such as recommender systems or reinforcement learning.

REFERENCES

[1] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1675–1684.

[2] X. Zhang, S. Tan, P. Koch, Y. Lou, U. Chajewska, and R. Caruana, "Axiomatic interpretability for multiclass additive models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 226–234.

[3] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. El-hadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. Sydney, NSW, Australia: ACM Press, 2015, pp. 1721–1730.

[4] X. Zhang, S. Tan, P. Koch, Y. Lou, U. Chajewska, and R. Caruana, "Axiomatic Interpretability for Multiclass Additive Models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19*. Anchorage, AK, USA: ACM Press, 2019, pp. 226–234.

[5] P. J. Darwen, "Bayesian model averaging for river flow prediction," *Applied Intelligence*, vol. 49, no. 1, pp. 103–111, 2019.

[6] B. Letham, C. Rudin, T. H. McCormick, D. Madigan *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[7] T. Wang, "Multi-Value Rule Sets," *arXiv:1710.05257 [cs], NIPS*, Oct. 2017, arXiv: 1710.05257.

[8] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, Sep. 2015, arXiv: 1511.01644.

[9] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," *arXiv preprint arXiv:1908.07442*, 2019.

[10] J.-S. Jang and C.-T. Sun, "Functional equivalence between radial basis function networks and fuzzy inference systems," *IEEE transactions on Neural Networks*, vol. 4, no. 1, pp. 156–159, 1993.

[11] S. Guillaume, "Designing fuzzy inference systems from data: An interpretability-oriented review," *IEEE Transactions on fuzzy systems*, vol. 9, no. 3, pp. 426–443, 2001.

[12] J.-S. Wang and C. G. Lee, "Self-adaptive neuro-fuzzy inference systems for classification applications," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 6, pp. 790–802, 2002.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[15] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," *arXiv preprint arXiv:1805.10820*, 2018.

[16] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1885–1894.

[17] P. Schwab, D. Miladinovic, and W. Karlen, "Granger-Causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4846–4853, Jul. 2019.

[18] P. Schwab and W. Karlen, "CXPlain: Causal Explanations for Model Interpretation under Uncertainty," *arXiv:1910.12336 [cs, stat]*, Oct. 2019, arXiv: 1910.12336.

[19] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 01 2009.

[20] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[21] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and causal parameters," *arXiv preprint arXiv:1608.00060*, 2016.

[22] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 4156–4165, 2019.

[23] M. Oprescu, V. Syrgkanis, and Z. S. Wu, "Orthogonal random forest for causal inference," *arXiv preprint arXiv:1806.03467*, 2018.

[24] D. J. Foster and V. Syrgkanis, "Orthogonal statistical learning," *arXiv preprint arXiv:1901.09036*, 2019.

[25] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, 1st ed. USA: Basic Books, Inc., 2018.

[26] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," *arXiv preprint arXiv:1902.02302*, 2019.

[27] T. Narendra, A. Sankaran, D. Vijaykeerthy, and S. Mani, "Explaining deep learning models using causal inference," *arXiv preprint arXiv:1811.04376*, 2018.

[28] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *ICML*, 2017.

[29] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining Classifiers with Causal Concept Effect (CaCE)," *arXiv:1907.07165 [cs, stat]*, Feb. 2020, arXiv: 1907.07165.

[30] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba, "Gan dissection: Visualizing and understanding generative adversarial networks," *arXiv preprint arXiv:1811.10597*, 2018.

[31] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf, "Counterfactuals uncover the modular structure of deep generative models," *arXiv preprint arXiv:1812.03253*, 2018.

[32] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," *arXiv preprint arXiv:1905.10958*, 2019.

[33] G. C. Cawley, "Causal & non-causal feature selection for ridge regression," in *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, ser. Proceedings of Machine Learning Research, I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, Eds., vol. 3. Hong Kong: PMLR, 03–04 Jun 2008, pp. 107–128.

[34] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, "Hiton: a novel markov blanket algorithm for optimal variable selection," in *AMIA annual symposium proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 21.

[35] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference using invariant prediction: identification and confidence intervals," *arXiv e-prints*, p. arXiv:1501.01332, Jan. 2015.

[36] R. M. Grath, L. Costabello, C. L. Van, P. Sweeney, F. Kamiab, Z. Shen, and F. Lecue, "Interpretable Credit Application Predictions With Counterfactual Explanations," *arXiv:1811.05245 [cs]*, Nov. 2018, arXiv: 1811.05245.

[37] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *arXiv:1711.00399 [cs]*, Mar. 2018, arXiv: 1711.00399.

[38] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Advances in neural information processing systems*, 2018, pp. 592–603.

[39] A. Van Looveren and J. Klaise, "Interpretable Counterfactual Explanations Guided by Prototypes," *arXiv:1907.02584 [cs, stat]*, Feb. 2020, arXiv: 1907.02584.

[40] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM, Jan. 2020, pp. 607–617.

[41] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[42] C. Russell, "Efficient Search for Diverse Coherent Explanations," in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. Atlanta, GA, USA: ACM Press, 2019, pp. 20–28.

[43] A. Artelt and B. Hammer, "Convex density constraints for computing plausible counterfactual explanations," *arXiv preprint arXiv:2002.04862*, 2020.

[44] S. Sharma, J. Henderson, and J. Ghosh, "Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models," *arXiv preprint arXiv:1905.07857*, 2019.

[45] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 344–350.

[47] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[48] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," *Journal of Business & Economic Statistics*, vol. 0, no. 0, pp. 1–10, 2019. [Online]. Available: https://doi.org/10.1080/07350015.2019.1624293

[49] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[50] J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan, "Axis: Generating explanations at scale with learnersourcing and machine learning," in *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, ser. L@S '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 379–388. [Online]. Available: https://doi.org/10.1145/2876034.2876042

[51] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[52] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vision*, vol. 45, no. 2, p. 83–105, Nov. 2001. [Online]. Available: https://doi.org/10.1023/A:1012460413855

[53] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.

[54] M. Harradon, J. Druce, and B. Ruttenberg, "Causal learning and explanation of deep neural networks via autoencoded activations," *arXiv preprint arXiv:1802.00541*, 2018.

[55] "DoWhy: A Python package for causal inference," https://github.com/microsoft/dowhy.

[56] R. H. Dehejia and S. Wahba, "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and statistics*, vol. 84, no. 1, pp. 151–161, 2002.

[57] C. E. Frangakis and D. B. Rubin, "Principal stratification in causal inference," *Biometrics*, vol. 58, no. 1, pp. 21–29, 2002.

[58] M. Research, "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation," https://github.com/microsoft/EconML, 2019, version 0.x.

[59] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao, "Causalml: Python package for causal machine learning," 2020.

[60] Y. Zhao, X. Fang, and D. Simchi-Levi, "Uplift Modeling with Multiple Treatments and General Response Types," *arXiv e-prints*, p. arXiv:1705.08492, May 2017.

[61] N. J. Radcliffe and P. D. Surry, "Real-world uplift modelling with significance-based uplift trees," *White Paper TR-2011-1, Stochastic Solutions*, pp. 1–33, 2011.

[62] D. Kalainathan and O. Goudet, "Causal Discovery Toolbox: Uncover causal relationships in Python," *arXiv e-prints*, p. arXiv:1903.02278, Mar. 2019.

[63] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, 2019. [Online]. Available: https://advances.sciencemag.org/content/5/11/eaau4996

[64] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018. [Online]. Available: https://doi.org/10.1063/1.5025050

[65] ——, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," *arXiv preprint arXiv:1709.01447*, 2017.

[66] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

[67] J. Runge, "Quantifying information transfer and mediation along causal pathways in complex systems," *Physical Review E*, vol. 92, no. 6, p. 062829, 2015.

[68] Y. Xian, Z. Fu, S. Muthukrishnan, G. De Melo, and Y. Zhang, "Reinforcement knowledge graph reasoning for explainable recommendation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 285–294.