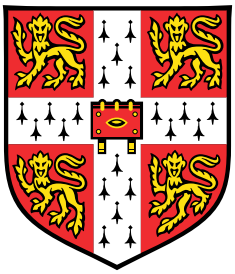
# Interpretability for Conditional Average Treatment Effect Estimation



**Javier Abad Martínez**

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
*Master of Philosophy in Machine Learning and Machine Intelligence*

Clare College August 2021

# Declaration

I, Javier Abad Martínez of Clare College, being a candidate for the MPhil in Machine Learning and Machine Intelligence, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

**Software** All computing experiments were carried out in Python. The CATE estimators were implemented from scratch in PyTorch and rely on standard libraries. Regarding the interpretability methods, Integrated Gradients was implemented from scratch, the Captum library was used for LIME and DeepLIFT, and the SHAP library for SHAP. No other third-party software was used. The full experimentation is included at

**/home/ja718/rds/hpc-work/Project**.

**Word Count** 14,991

Javier Abad Martínez
August 2021

# Acknowledgements

First and foremost, I would like to thank my supervisor, Mihaela van der Schaar, for introducing me to the topic and for her invaluable guidance throughout the thesis. I am also enormously grateful to have been advised by Jonathan Crabbé and Alicia Curth. Thank you for being generous with your time and for all the enlightening conversations that have been vital for the development of the thesis. Special appreciation goes to Rupert Brown for proofreading.

I cannot forget to thank my Canary Islands Team. Thank you for being amazing friends, meeting you has been the highlight of the year.

Finally, I would like to dedicate this thesis to my loving parents. You are the reason behind my successes and the inspiration behind my endeavors. Thank you for all you have done and all that you are still doing.

# Abstract

Machine learning has recently shown much potential for causal inference tasks in several important fields. Of particular interest is the estimation of individual-level treatment effects in the context of precision medicine, which can provide clinicians with crucial information for the design and issuance of patient-specific treatment plans. This implies answering questions of the type: "Will chemotherapy increase the life expectancy of this oncologic patient?" However, to effectively integrate machine learning models in the medical setting, they must also be interpretable, so that clinicians can understand and validate their decisions. Surprisingly, little attention has been paid to the intersection of the two fields – individual treatment effect estimation and interpretability – and the only works that explicitly consider it frame the problem as interpreting a standard supervised learning setting, ignoring the idiosyncrasies of the treatment effect estimation problem.

The present work proposes a framework for interpreting the Conditional Average Treatment Effect (CATE) estimation problem. By examining the unique features of the CATE setting, our method allows identification of the information that is of the utmost importance in the design of personalized policies: the predictive covariates. Furthermore, the solid theoretical formalization of the framework enables an extensive characterization of patients at the individual level, for instance, by additionally disentangling the prognostic and irrelevant covariates and uncovering the specific features that make a treatment more or less efficacious for an individual compared to the average population.

The proposed framework is extensively evaluated in both synthetic and real-world experiments. We present use-cases to provide practitioners with specific guidelines on using the framework, for example, by identifying which covariates determine the increase in cognitive test scores for premature children that join a health development program. Finally, we characterize four popular CATE estimators based on their ability to disentangle the predictive covariates. We do so in several synthetic datasets inspired by the drug development setting, thus providing insights on design choices by outlining the estimators' virtues and limitations in various scenarios. We therefore show how our method can serve as a tool for model selection, which is naturally challenging in causal inference tasks.

# Table of contents

# List of figures

# List of tables

# Nomenclature

**Acronyms / Abbreviations**

| | |
|---|---|
| AIPW | Augmented Inverse Probability Weighting |
| ASI | Affine Scale Invariance |
| ATE | Average Treatment Effect |
| AUROC | Area Under the Receiver Operating Characteristic curve |
| CATE | Conditional Average Treatment Effect |
| DeepLIFT | Deep Learning Important FeaTures |
| DGP | Data Generation Process |
| DL | Deep Learning |
| EHR | Electronic Health Record |
| ELU | Exponential Linear Unit |
| FPR | False Positive Rate |
| IG | Integrated Gradients |
| IHDP | Infant Health and Development Program |
| IPM | Integral Probability Metric |
| ITE | Individualized Treatment Effect |
| LIME | Local Interpretable Model-agnostic Explanations |
| ML | Machine Learning |
| MMD | Maximum Mean Discrepancy |

| | |
|---|---|
| MSE | Mean Squared Error |
| NN | Neural Network |
| PEHE | Precision in Estimation of Heterogeneous Effect |
| PO | Potential Outcome |
| RCT | Randomized Control Trial |
| RKHS | Reproducing Kernel Hilbert Space |
| ROC | Receiver Operating Characteristic |
| SHAP | SHapley Additive exPlanations |
| SUTVA | Stable Unit Treatment Value Assumption |
| TE | Treatment Effect |
| TPR | True Positive Rate |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |

# Chapter 1

# Introduction

Inferring the causal effect of interventions is a fundamental problem in many domains, including economics, education, and healthcare. Much recent work focuses on the last of these, particularly on assisting in the design of personalized plans in the context of precision medicine. For instance, a clinician might be interested in which medication will cause better outcomes for a specific patient. Deciding on personalized interventions of this kind involves tailoring treatments to patients, which requires estimating individual-level causal effects based on the information at hand. This can be framed as estimating the Conditional Average Treatment Effect (CATE)[1].

In medicine, prospective experiments like Randomized Control Trials (RCTs) are the *de facto* gold standard for causal inference, allowing evaluation of the treatment effectiveness across a population by randomly allocating it. However, fully randomized studies are often infeasible, prohibitively expensive, or even face some ethical issues (Schafer, 1982). Furthermore, their focus on specific questions usually does not provide a complete characterization of the population heterogeneity, which makes them unsuitable for estimating individual-level treatment effects. Real-world observational datasets have become a tempting shortcut, mainly due to the widespread accumulation of data in the form of Electronic Health Records (EHRs). EHRs contain large amounts of clinical information about heterogeneous patients and their responses to treatments, which data-driven machine learning (ML) methods can use to support clinical decision-making (Bica et al., 2021). Using ML for estimating individual-level treatment effects has become a thriving area of research and has been proven to provide clinicians with actionable intelligence for making treatment decisions, such as clinical risk assessment (Alaa and van der Schaar, 2018c; McCauley and Darbar, 2016) or treatment response prediction (Athreya et al., 2019).

[1] sometimes also referred to as Individualized Treatment Effect (ITE).

In general, the more data available, the more accurate estimation an ML algorithm can make. However, the simple pursuit of predictive accuracy is insufficient in clinical practice, where understanding the model should also be guaranteed. With the increasing availability of EHRs, Neural Networks (NNs) and other Deep Learning (DL) methods have achieved great success in the healthcare domain, including in CATE estimation. However, typical DL models are not interpretable – meaning that they do not inherently explain their predictions – which impedes their adoption in clinical decision-making. For example, in a study funded by Cost-Effective Healthcare, a DL model was considered too risky to predict the mortality of patients with pneumonia because it could not be understood, despite its remarkable accuracy (Caruana et al., 2015; Cooper et al., 2005). Therefore, to adequately integrate these ML algorithms in the medical setting, clinicians must be able to understand and trust their decisions. In other words, they must be able to interpret the models.

There has been much progress made independently in both the CATE and interpretability fields, but the literature at their intersection is sparse at best. Moreover, the only studies that explicitly consider interpretability in the CATE context frame the problem as explaining a standard supervised learning setting. Nonetheless, we argue that an "A+B solution" is insufficient, and that the unique features of the CATE estimation problem should be considered in the context of interpreting models.

The present work builds on the conviction that, to interpret CATE estimation effectively, the idiosyncrasies of the problem must be taken into account. Under this premise, we develop a framework where we unify both fields, and we give guidelines on how to employ it from the perspective of a clinical practitioner. We additionally show the potential of the framework in an extensive qualitative and quantitative evaluation in both synthetic and real-world experiments.

## 1.1 Thesis Contributions

The main contributions of this thesis are as follows:

1. A review of existing machine learning methods for CATE estimation, particularly of a family of non-parametric regression approaches. The various methods are gathered in a taxonomy that characterizes them by their strategies at estimating individual-level treatment effects.

2. A re-axiomatization of the properties of four well-known saliency methods in the context of interpreting CATE. We outline which saliency methods satisfy which properties, and we provide guidelines on how they can be employed to support clinical decision-making, all this in a unified notation that links to the CATE setting.

3. A novel framework for interpreting CATE estimation, highlighting how the problem differs from the standard supervised learning paradigm and previous research on the topic. We demonstrate how saliency methods can be utilized in this context, given their interpretability mechanisms and the outlined properties, and we evaluate them quantitatively and qualitatively in both synthetic and real-world experiments.

4. A comprehensive comparison of four popular CATE estimators based on their ability to discover predictive covariates – which we relate to the problem of interpreting CATE. We characterize the estimators in a range of synthetic datasets inspired by the drug development setting.

## 1.2 Thesis Outline

The structure of this thesis is as follows:

**Chapter 2** provides the theoretical background necessary to understand the methods and ideas presented in the rest of the thesis. We cover the problem of estimating CATE, and we give a brief introduction to the field of interpretability. We then review the sparse literature on interpretability for CATE estimation, highlighting its weaknesses and misspecifications and how our framework differs from it.

**Chapter 3** presents four well-known saliency methods employed in traditional machine learning interpretability. We describe them in a unified notation that links to the CATE setting.

**Chapter 4** introduces the proposed framework. We first formalize the problem of interpreting CATE and outline how the saliency methods from Chapter 3 can help accomplish our goal. We next describe four CATE estimators, later compared in Chapter 5, and we propose an extension to their original architectures, which we justify by empirical evaluations. Finally, we further illustrate how saliency methods are practical for interpreting CATE by identifying a set of desirable properties that can support clinical decision-making.

**Chapter 5** extensively evaluates the proposed framework. We first set a CATE estimator and compare the saliency methods quantitatively and qualitatively, additionally giving guidelines on how they can be used to assist in the clinical setting, both in synthetic and real-world experiments. We then fix a saliency method and evaluate the CATE estimators considered in Chapter 4 based on their ability to disentangle the predictive features. We do so in a range of synthetic datasets inspired by the drug development setting.

**Chapter 6** provides some final thoughts and directions for future work.

# Chapter 2

# Background Theory

This chapter presents the theoretical underpinnings of the proposed method. Our framework attempts to intersect two vast machine learning fields: treatment effect estimation and interpretability. Section 2.1 begins with an introduction to the CATE setting and to its main challenges, leading to a description of a common framework to conceptualize the problem and a taxonomy that contextualizes the considered models. Next, Section 2.2 introduces the reader to the field of interpretability. We go through some of its definitions, its taxonomy, and its relevance in the domain of healthcare. Finally, Section 2.3 reviews the sparse literature in the intersection of interpretability and CATE. We identify the limitations and misspecifications of the current approaches, which further motivate the development of our framework.

## 2.1 The CATE estimation problem

The Treatment Effect (TE) is the difference between outcomes under different treatments. The present work focuses on drawing inferences about individual-level TEs, as opposed to simply estimating the Average Treatment Effect (ATE) for the whole population. This is of the utmost importance in the domain of healthcare, where a clinician might be interested in how prescribing a medication will influence or determine an outcome for a particular patient. In this context, treatment guidelines based on the average patient are insufficient, and the need to account for the specific features of the individual arises. This can be achieved by estimating the Conditional Average Treatment Effect (CATE).

This section first reviews the challenges of the CATE estimation problem, and presents a standard framework to conceptualize it. We next explain how CATE can be computed using machine learning, and we show how different approaches relate to each other in a proposed taxonomy.

### 2.1.1 The challenges of CATE estimation

The CATE estimation involves computing the treatment effect for a specific individual. In this sense, we might want to predict the increase in the life expectancy of an oncologic patient after administering chemotherapy, given by the difference between the individual's life expectancy under the two possible scenarios, namely the treated and the untreated. However, the available EHRs used to fit the ML-based estimators usually contain just the observed outcomes. The fact that we do not have access to the counterfactual outcomes[1] is often referred to in the causality literature as *the fundamental problem of causal inference*, and it is one of the reasons why the CATE estimation problem is fundamentally different from the standard supervised learning paradigm (Alaa and van der Schaar, 2018b; Imbens and Rubin, 2015).

Another inherent challenge of learning from observational data is the existence of confounders. Confounders are variables that determine both the treatment assignment and the outcome. For example, in the clinical setting the patient's age determines the treatment applied – e.g., younger patients receiving surgery while older patients painkillers – and the outcome itself – e.g., younger patients showing a better recovery rate. Confounders can introduce a spurious, non-causal behavior in the treatment effect (Johansson et al., 2020). For instance, if we compute the treatment effect simply by averaging the recovery rate across the "painkillers" group and the "surgery" group, we might conclude that surgery is a more effective treatment because most of the patients in that group are young patients. Table 2.1 exemplifies this phenomenon with the *Simpson's paradox* (Julious and Mullee, 1994). Here, painkillers have a higher recovery rate for both young and old patients, but when combining the groups surgery turns out to be the better treatment. The effect of the age is mistakenly counted into the effect of the treatment on the outcome, thus generating a spurious effect.

| | **Surgery** | **Painkillers** |
|---|---|---|
| **Old** | $25/40 = 62.5\%$ | $120/180 = \mathbf{66.7}\%$ |
| **Young** | $140/160 = 87.5\%$ | $18/20 = \mathbf{90}\%$ |
| **Total** | $165/200 = \mathbf{82.5}\%$ | $138/200 = 69\%$ |

**Table 2.1** Example of a spurious behavior generated by a confounder with the *Simpson's paradox* (Julious and Mullee, 1994). Here, age is a confounder affecting the outcome (recovery rate) and the treatment assignment (surgery or painkillers). If we estimate the $\hat{\text{ATE}}$: 165/200-138/200=13.5%, it erroneously draws that surgery is more effective than painkillers. This is because most patient of the "surgery" group are young (160/200) compared to the "painkillers" group (20/200), and young people have a higher recovery rate regardless of treatment.

Moreover, the existence of confounders means that treatments are not assigned at random. This is mostly due to the clinicians introducing bias into the data from their decision-making

[1] Counterfactual outcomes answer questions of the type "What would have happened if the patient had (not) been treated?"

process, e.g., by administering different treatments depending on the patient's age. This selection bias implies having unbalanced patient representativeness in different regions of the covariate space, thus decreasing the accuracy of predicting the counterfactuals in those regions. In other words, selection bias implies different covariate distributions in the treatment groups, resulting in the estimator performing poorly for the "painkillers" group if trained in the "surgery" group and vice versa. Figure 2.1 illustrates this phenomenon.

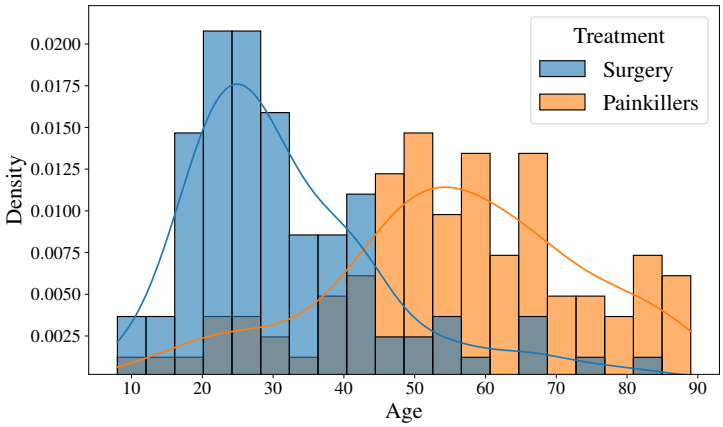



**Fig. 2.1** Example of selection bias induced by the confounder "Age". We illustrate the covariate shift between the two treatment distributions, "Painkillers" and "Surgery".

Several approaches have been proposed to frame and solve the challenges above. We now introduce one of the most popular ones: the Neyman-Rubin framework.

### 2.1.2 The Neyman-Rubin Potential Outcomes framework

The present work conceptualizes the CATE estimation problem using the Neyman-Rubin potential outcomes framework (Rubin, 2005), also known as the Rubin Causal Model.

We assume an observational dataset $\mathcal{D} = \{(Y_i, \boldsymbol{X}_i, W_i)\}_{i=1}^{n}$ with $(Y_i, \boldsymbol{X}_i, W_i) \overset{i.i.d.}{\sim} \mathcal{P}$, where $Y \in \mathcal{Y}$ is a binary or continuous outcome of interest (e.g., one-year risk of stroke), $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector with the patient's covariates (e.g., age, wealth status, cholesterol level), and $W_i \in \{0, 1\}$ is a binary treatment (e.g., {0: placebo, 1: drug}) assigned according to a propensity score $\pi(\boldsymbol{x}) = \mathbb{P}(W = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. The two potential outcomes (POs) represent the outcome under the application of the treatment $Y_i^{(1)}$ and when the treatment is not applied $Y_i^{(0)}$ (Figure 2.2).

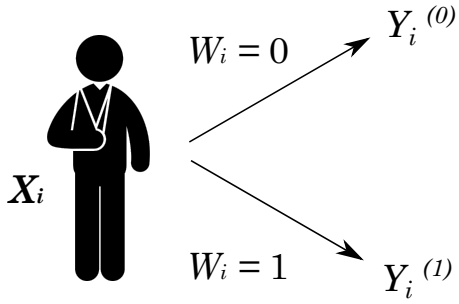



**Fig. 2.2** Illustration of the Neyman-Rubin potential outcomes framework. A patient $\boldsymbol{X}_i$ receives a treatment $W_i = w$ with $w \in \{0, 1\}$, and we observe only the (factual) outcome $Y_i^{(w)}$.

Due to the aforementioned *fundamental problem of causal inference*, we only have access to the factual (i.e., observed) outcome, so that

$$Y_i = W_i Y_i^{(1)} + (1 - W_i)\, Y_i^{(0)}. \tag{2.1}$$

The CATE is defined as the expected difference between the two POs conditioned on the covariates $\boldsymbol{X} = \boldsymbol{x}$[2]:

$$\tau(\boldsymbol{x}) := \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid \boldsymbol{X} = \boldsymbol{x}\right]. \tag{2.2}$$

We say that the CATE is *identifiable* if it is possible to obtain a consistent estimate of Equation (2.2) from data (Imbens and Wooldridge, 2009; Pearl et al., 2009). For this to happen, the following assumptions[3] must hold:

**Assumption 1.** *Consistency. If an individual is assigned treatment $w$, we observe the associated potential outcome $Y = Y(w)$.*

**Assumption 2.** *Stable Unit Treatment Value Assumption (SUTVA). The potential outcomes for any individual are independent of the treatment assigned to other individuals, and there are no other versions of this treatment that lead to different potential outcomes.*

**Assumption 3.** *Ignorability. There are no hidden confounders, the potential outcomes are independent of the treatment assignment given the covariates, i.e., $Y^{(0)}, Y^{(1)} \perp W \mid \boldsymbol{X}$.*

**Assumption 4.** *Positivity. The treatment assignment is not deterministic, that is, each individual has a non-zero probability of belonging both to the control and the treated groups, i.e., $0 < \pi(\boldsymbol{x}) < 1, \forall \boldsymbol{x} \in \mathcal{X}$.*

[2] From now on we drop index $i$ for patient.

[3] Note that the validity of the ignorability assumption cannot be assessed from data, and must be determined by domain knowledge and understanding of the causal relationships between the variables.

Given *identifiability*, Equation (2.2) can be rewritten as:

$$\begin{aligned}
\tau(\boldsymbol{x}) &= \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[Y^{(1)} \mid \boldsymbol{X} = \boldsymbol{x}\right] - \mathbb{E}\left[Y^{(0)} \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[Y^{(1)} \mid \boldsymbol{X} = \boldsymbol{x}, W = 1\right] - \mathbb{E}\left[Y^{(0)} \mid \boldsymbol{X} = \boldsymbol{x}, W = 0\right] \qquad (2.3) \\
&= \mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}, W = 1\right] - \mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}, W = 0\right] \qquad (2.4) \\
&= \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}), \qquad (2.5)
\end{aligned}$$

where Equality (2.3) is due to the ignorability assumption, and Equality (2.4) follows from the consistency and the SUTVA assumptions. Equality (2.5) shows that the CATE can be expressed with the difference between the response surface under treatment, $\mu_1(\boldsymbol{x})$, and under control, $\mu_0(\boldsymbol{x})$, defined as

$$\mu_w(\boldsymbol{x}) := \mathbb{E}\left[Y^{(w)} \mid \boldsymbol{X} = \boldsymbol{x}\right]. \qquad (2.6)$$

Note that the positivity assumption guarantees that, for each patient $\boldsymbol{X} = \boldsymbol{x}$, it is possible to observe samples from the two $Y^{(w)}$ – although in practice we only observe one – and, therefore, that we can estimate the surfaces $\mu_w$ from data.

In conclusion, if assumptions 1-4 hold, the CATE can be calculated using standard machine learning algorithms trained with observational data under the Neyman-Rubin framework. A plethora of methods has been proposed in the literature to this end. The present work focuses on the so-called meta-learners, a family of non-parametric regression models introduced below.

### 2.1.3 Meta-learners for CATE estimation

Formally, non-parametric regression approaches model the outcome surface $Y$ as a function of the treatment assignment $W$, the covariates $\boldsymbol{X}$, and some unobservable noise $\epsilon$, which we assume to be additive Gaussian. This leads to the expression:

$$Y = f(\boldsymbol{X}, W) + \varepsilon, \quad \text{with} \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right), \qquad (2.7)$$

where $f(\boldsymbol{X}, W) = \mathbb{E}[Y \mid \boldsymbol{X}, W]$. Given this framework, several meta-learners have been proposed to estimate CATE.

First introduced by Kunzel et al. (Künzel et al., 2019), these meta-learners allow decomposition of the problem of estimating CATE into regression tasks that can be solved using standard supervised learning techniques, such as tree ensembles (Athey and Imbens, 2016; Lu et al., 2018; Powers et al., 2018) and Neural Networks (Johansson et al., 2020; Shalit

et al., 2017). Here we follow the taxonomy proposed in (Curth and van der Schaar, 2021b), differentiating between *direct* and *indirect* meta-learners.

### Direct meta-learners

A *direct* meta-learner targets $\tau(\boldsymbol{x})$ directly in a two-step fashion, first obtaining some *nuisance* parameters $\eta = (\mu_0(\boldsymbol{x}), \mu_1(\boldsymbol{x}), \pi(\boldsymbol{x}))$ from observational data and then estimating $\hat{\tau}(\boldsymbol{x})$ by regressing a pseudo-outcome $\tilde{Y}_{\hat{\eta}}$ (based on $\hat{\eta}$) on the covariates $\boldsymbol{X}$. For pseudo-outcome $\tilde{Y}_{\hat{\eta}}$ it holds that $\mathbb{E}\left[\tilde{Y}_{\eta} \mid \boldsymbol{X} = \boldsymbol{x}\right] = \tau(\boldsymbol{x})$, that is, it is an unbiased CATE estimator if $\eta$ is known.

Different strategies estimate different *nuisance* parameters in the first step, including the propensity score $\pi(\boldsymbol{x})$ and/or the POs response surfaces $\mu_w(\boldsymbol{x})$. Kennedy's DR-learner (Kennedy, 2020) first obtains separately $\hat{\pi}(\boldsymbol{x})$ and the two response surfaces $\hat{\mu}_w(\boldsymbol{x})$, and then regresses a pseudo-outcome that is unbiased if any of the them is correctly predicted. The X-learner (Künzel et al., 2019) computes two group-specific estimators, $\hat{\tau}_1(\boldsymbol{x})$ and $\hat{\tau}_0(\boldsymbol{x})$, and combines them with a weighting function to estimate the global $\hat{\tau}(\boldsymbol{x})$ . In (Curth and van der Schaar, 2021a), the authors also introduce two simplified versions of the DR-learner, namely the RA-learner and the PW-learner, which exclusively estimate the response surfaces and the propensity score in their first step, respectively. We provide details on the DR- and the X-learner in Section 4.2.

### Indirect meta-learners

Unlike previous work (e.g., (Künzel et al., 2019)), we define as *indirect* meta-learner *any* estimator that first fits the POs response surfaces $\hat{\mu}_w(\boldsymbol{x})$ and then computes CATE by their difference $\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x})$. This way, we consider a broader classification than that underlying Kunzel's S- and T-learners. Notably, we include multi-task learning methods, representation learning methods, and disentangled representation methods; which are traditionally classified in separate categories, e.g., in (Caron et al., 2020; Yao et al., 2020).

**Multi-task learning methods** conceptualize CATE estimation as a multi-task learning problem. They use a neural architecture with shared layers for the treated and control groups, on top of which private layers are set for each group separately. There are several exciting extensions to the multi-task paradigm based on the Bayesian framework, for instance, in the form of multi-task Gaussian processes (Alaa and van der Schaar, 2017). The probabilistic treatment of the problem allows accounting for uncertainty, providing measures of confidence for the POs and CATE estimates, which are of the utmost importance in the context of precision medicine. We also include (Alaa et al., 2017)'s DCN-PD in this category, which employs Monte-Carlo propensity-dropout at inference, alleviating the selection bias problem and tracking individualized uncertainty measures.

**Representation learning methods** (Bengio et al., 2013; Shalit et al., 2017) formally involve jointly learning a representation $\Phi : \mathcal{X} \to \mathcal{S}$ and two regression heads $h_w : \mathcal{S} \to \mathcal{Y}$, each fit with the corresponding treatment/control group data in the latent space. For instance, Shalit's TARNet (Shalit et al., 2017) minimizes a weighted sum of the factual loss by training this neural architecture in an end-to-end fashion. DragonNet (Shi et al., 2019) additionally predicts the propensity score to enforce the representation space to keep covariates relevant for treatment assignment, and learns the hypothesis of both groups on top of it. Specifically, representation learning methods have become a popular approach to combat the selection bias challenge. Recall that selection bias implies having different covariate distributions in the treated and control groups. This covariate shift can be framed as a domain adaptation problem (Mansour et al., 2009). In this sense, some methods attempt to enforce a balanced representation where the two distributions look similar by minimizing an Integral Probability Metric (IPM) measure of distance between them. This is the case of the CFR (Shalit et al., 2017), the RCFR (Johansson et al., 2018), and ACE (Yao et al., 2019).

However, the aforementioned representation learning methods do not differentiate between covariates that are purely predictive of outcome $Y$, treatment $W$, or both (i.e., confounders). Therefore, balancing all covariates in the representation space could counterproductively induce additional bias since some of them determine the treatment assignment. **Disentangled representation methods** attempt to solve this problem by learning to decompose the representations of confounders and non-cofounders. Therefore, they allow minimizing exclusively the distance between the distribution of covariates that only affect the outcome $Y$, so that we do not discard information about treatment assignment. Prominent examples of disentangled representation methods are Hassanpour et al.'s DR-CFR (Hassanpour and Greiner, 2019b) and Wu et al.'s DeR-CFR (Wu et al., 2020). Additionally, Curth and VdS's SNet (Curth and van der Schaar, 2021a) further adds two extra representations for predictive covariates that affect a single potential outcome, namely $Y^{(0)}$ and $Y^{(1)}$, thus accounting for the heterogeneity of treatment effectiveness across groups.

It is easy to spot similarities between the indirect CATE estimation approaches (Figure 2.3). The reality is that they are many times complementary, and a single meta-learner could use techniques from the three domains (multi-task learning, representation learning, and disentangled representation learning). In fact, SNet has been shown to generalize some of the models mentioned above, including TARNet, DragonNet, DeR-CFR, and TNet (i.e., NN-based T-learner). Section 4.2.4 revisits these features in detail, pointing out their relevance to our specific problem.

This section has detailed the problem and challenges of estimating individual-level treatment effects. It has also reviewed several approaches to estimate CATE accurately, presenting a family of non-parametric regression models. However, if the goal is designing treatment policies for personalized medicine, we argue that it is at least equally important to

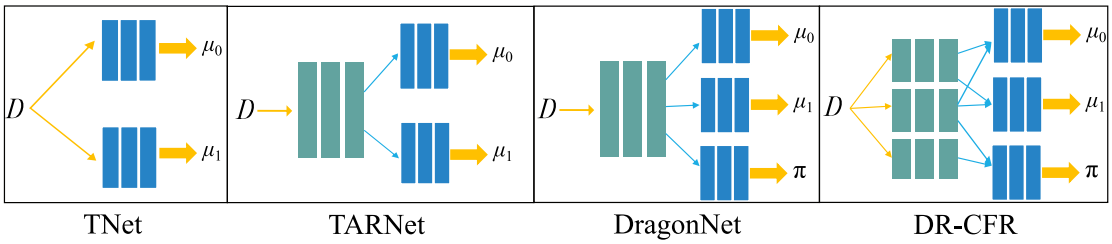



**Fig. 2.3** Model architectures of four well-known indirect CATE estimators. TNet simply fits each POs surface with a NN. TARNet, DragonNet and DR-CFR first use a shared representation and estimate the POs and/or the propensity score with private layers. DR-CFR considers a disentangled representation of the shared layers by decomposing them in three representations. TARNet, DragonNet and DR-CFR clearly follow a multi-task learning approach. All estimators compute CATE by the difference of the estimated POs surfaces, which further supports that the four should be considered in the same category of *indirect* meta-learners. Adapted from (Curth and van der Schaar, 2021a)

| Category | Strategy | Examples |
|---|---|---|
| **Direct** | Estimates $\mu_w$ | DR-learner, RA-learner |
| | Estimates $\pi$ | DR-learner, PW-learner |
| | Other | X-learner |
| **Indirect** | T-learners | TNet |
| | Multi-task learning | DCN-PD, TARNet, DragonNet, CFR, ACE, DR-CFR, SNet |
| | Representation learning | TARNet, DragonNet, CFR, ACE, DR-CFR, DeR-CFR, SNet |
| | Disentangled representation learning | DR-CFR, DeR-CFR, SNet |

**Table 2.2** Summary of some meta-learner approaches. Note that one estimator can belong to different categories.

assess whether an algorithm leads to the correct *interpretation* of the drivers of the underlying treatment effect heterogeneity. The following section introduces the reader to the field of interpretability, describing a set of methods that could shed light on this matter and laying the foundations for our proposed framework.

## 2.2 Machine Learning Interpretability

ML has been successfully deployed in a wide range of domains, including healthcare, finance, autonomous driving, and speech recognition. A clear example is the family of meta-learners described in Section 2.1.3, with a direct application in clinical decision-making, drug discovery, and precision medicine. This often means that decisions that humans used to make are now made automatically by these algorithms.

In order to ensure the reliability of such decisions, humans need to understand *how* and *why* they are made (Moraffah et al., 2020). However, ML models are generally *black-boxes*, meaning that they do not inherently explain their decisions. This is particularly problematic given the impending regulation on the need for explanatory systems like the European Union's "Right to Explanation" (Goodman and Flaxman, 2017), which states that an applicant has

the so-called right to be informed and could require a list of all the factors that influence a model's decision. Moreover, research on the topic has uncovered racially-biased ML systems applied to criminal risk assessment (Angwin et al., 2016) and evidenced the fallibility of Deep Neural Networks when minimally perturbing their inputs (Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2016). Therefore, understanding the decisions of ML models can help identify their weaknesses and sources of bias, which is a path towards increasing their safety. The field that researches in this direction is the *Machine Learning Interpretability*.

This section briefly introduces the reader to the field of interpretability. We review some of its definitions, as well as a typical taxonomy. Finally, we justify the importance of interpretability in the domain of healthcare, and, more specifically, in the context of precision medicine. Note that this section does not aspire to be an exhaustive review of the interpretability field. Conversely, the goal here is to give some background to contextualize our problem and to establish the basis for our proposed method.

### 2.2.1 Definition

There has been a growing interest[4] in the interpretability field due to the impressive performance of non-interpretable models in high-stakes decision-making, where not explaining the algorithm's reasoning presents evident dangers (Adadi and Berrada, 2018). However, despite the apparent relevance and enthusiasm, there is no consensus among the research community on *what* we mean by interpretability. Notably, Miller (Miller, 2019) defines it as "the degree to which a human can understand a cause of a decision." Kim et al. (Kim et al., 2016), on the other hand, suggest that interpretability is the degree to which humans can consistently predict the model's decision. Doshi-Velez et al. (Doshi-Velez and Kim, 2017) understand interpretability as the ability to present information in intelligible terms.

In the present work, we are more aligned with Gilpin's (Gilpin et al., 2018) goal-oriented conception of interpretability: "The goal of interpretability is to describe the internals of a system in a way that is understandable to humans." In this sense, by *interpreting* a model here we mean answering the question: "Why does this particular input lead to that particular output?" or, framing in terms of our problem: "Why does this particular patient lead to that particular treatment effect?".

### 2.2.2 Taxonomy

We follow the taxonomy of *traditional* interpretable models introduced in (Moraffah et al., 2020) and divide these algorithms into two categories: (1) inherently interpretable models, and (2) post-hoc model interpretability.

[4] +40,000 related publications since 2015 according to Google Scholar.

**Inherently interpretable models** These models either embed explanations as part of their architecture or are constrained to enhance their simplicity, hence their intelligibility. Notable interpretable models are the decision trees, whose predictions can be explained by tracing back the conditions on their nodes; rule-based models, which infer predictions by understandable *if... then...* rules; linear regressions, whose weights have a straightforward connection with feature importance; and disentangled representations, which break down features into independent latent variables that are highly correlated with meaningful patterns (e.g., PCA (Jolliffe, 2005) or deep latent-variable models such as VAE (Kingma and Welling, 2013)).

**Post-hoc interpretability** Post-hoc interpretability aims to explain the decisions of a black-box model after training. A prominent family of post-hoc interpretability methods is saliency methods, which explain a model's decision by highlighting the input features that are most relevant for issuing the prediction. According to (Crabbé and van der Schaar, 2021), saliency methods can be gradient-based (e.g., DeepLIFT (Shrikumar et al., 2017), Integrated Gradients (Sundararajan et al., 2017)), perturbation-based, attention-based, or have their own mechanism for computing feature importance (e.g., LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017)). Other post-hoc interpretability methods include example-based explanations and feature visualization, which generally deal with image inputs. Furthermore, black-box models can be explained by using base interpretable models, such as decision trees, rule-based models, or linear regression. In this sense, an inherently interpretable model is trained to approximate the behavior of a more complex black-box, like a NN (Boz, 2002).

There is a separate category of interpretability methods that has gained much interest recently: causal interpretability [5]. This framework, first introduced by Pearl (Pearl, 2018), outlines the problem of interpretability by understanding the causal-effect relations of a model, not just the correlations captured by traditional interpretability techniques. It is therefore not to be confused with our problem at hand, which is studying how (traditional) interpretability applies to CATE estimation, which is a causal inference problem by nature. Causal interpretability, on the other hand, conceives the problem of interpreting a model from a causality framework.

Having defined *what* we mean by interpretability and its general taxonomy, the following section examines its relevance in the context of our problem: the clinical setting.

[5] We refer the interested reader to two comprehensive surveys on causal interpretability: (Moraffah et al., 2020) and (Xu et al., 2020).

### 2.2.3 Interpretability in Healthcare

The healthcare domain is arguably one of the most challenging applications of machine learning. The medical field contemplates additional risks and responsibilities that other fields do not consider when incorporating ML systems. Notably, in clinical decision-making lives may be at stake, which raises ethical concerns about using these systems if they are not reliable (Jiang et al., 2017). Moreover, in clinical decision support we are confronted with unknown, incomplete, noisy, erroneous, and inaccurate datasets in arbitrarily high-dimensional spaces (Holzinger et al., 2014). This problem is even more critical in personalized medicine, since it requires the fusion of these various heterogeneous data sources to predict treatment effects at the individual level. We argue that to integrate machine learning algorithms in this context effectively, the clinician must have the possibility to understand *how* and *why* the model has made a decision. Simply put: the clinician must be able to *interpret* the model.

Interpretability can help understand how diverse and heterogeneous data may contribute to a *relevant* prediction at the patient level. In (Tonekaboni et al., 2019), the authors surveyed 10 clinicians with varying years of experience to identify the specific aspects of interpretability that could build trust in ML. They unanimously agreed that knowing the subset of features that defines the model's outcome is crucial in the clinician's decision-making, allowing them to compare the model's decision to their clinical judgment, especially in a discrepancy. Furthermore, they underlined the utility of feature importance measures for specific patients, as opposed to only at the population level.

These findings further motivate the present work, since our goal is to understand why a machine learning model recommends a particular treatment for a specific patient. Therefore, it seems sensible to tackle the problem by using interpretability methods that measure feature importance at the patient level. This will be the topic of Chapter 3. In the next section, we review the sparse literature on interpretability for CATE estimation, and we comment on how our approach differs from it.

## 2.3 Interpretability for CATE estimation

The ML literature at the intersection of CATE estimation and interpretability is currently sparse at best. Following the taxonomy in Section 2.2.2, a straightforward approach towards interpreting CATE uses inherently interpretable models for its estimation. Although simple, their implicit mechanisms for explaining the model's decisions have raised notable popularity among linear regression (Abrevaya et al., 2015; Hahn et al., 2018) and decision trees (Foster et al., 2011) for estimating treatment effects.

Nonetheless, such approaches can come at the cost of lower estimation accuracy compared to other methods. Notably, Neural Networks are the state–of–the–art in CATE estimation, principally by their representation learning approaches (Curth and van der Schaar, 2021a; Hassanpour and Greiner, 2019a; Shalit et al., 2017). NNs depend on an uninterpretable representation space, which implies that their decisions can only be explained by post-hoc interpretability.

The only work that explicitly considers the problem of interpreting CATE from the post-hoc perspective is (Kim and Bastani, 2019). The authors propose explaining a NN-based CATE estimator using a two-stage algorithm, fitting an inherently interpretable model on top of the already-trained POs regression surfaces. Specifically, they first learn an uninterpretable function $f^*$ to estimate the POs surfaces $h_w : \mathcal{S} \to \mathcal{Y}$ from the representation space $\Phi : \mathcal{X} \to \mathcal{S}$, which they define as the *oracle model* $f^*(\boldsymbol{X}, W) = h_W(\Phi(\boldsymbol{X}))$. In the second step, they use a supervised learning algorithm $\mathcal{A}$ to learn an interpretable model $\hat{f}$ (e.g., a linear regression) that approximates $f^*$. Formally:

$$\hat{f} = \mathcal{A}\left(\{(\boldsymbol{X}_i, W_i, f^*\left(\boldsymbol{X}_i, W_i\right))\}\right), \tag{2.8}$$

where $(\boldsymbol{X}_i, W_i) \sim p(\boldsymbol{X}, W)$. They choose $p(\boldsymbol{X}, W)$ to be the RCT distribution, where treatments $W$ are independent of the covariates $\boldsymbol{X}$ and are thus assigned randomly. Finally, they estimate CATE as $\hat{\tau}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}, 1) - \hat{f}(\boldsymbol{x}, 0)$.

However, we argue that the authors in (Kim and Bastani, 2019) do not consider the unique characteristics of CATE estimation in the context of interpreting models, and they frame the problem identically to explaining a standard supervised learning black-box. We believe that, unlike in a standard prediction setting, it is not most important to interpret how the patient's covariates influence the POs. Instead, we should focus on finding interpretations of the *difference* between (potential) outcomes under different treatments, which is the treatment effect itself – a feature that (Kim and Bastani, 2019) overlooks.

A way of characterizing this *difference* is by identifying which covariates affect both POs (and are thus discarded when computing the CATE) and which ones affect only one PO, accounting for treatment heterogeneity across groups. One therefore distinguishes between two types of covariates that determine the two components of patient responses, namely the prognostic and the predictive biomarkers (Ballman, 2015). We describe them next.

### 2.3.1 Disentangling predictive and prognostic biomarkers

The term *biomarker* relates to a measurement variable that is associated with a disease outcome (Ballman, 2015). We will use the word *biomarker* interchangeably with *covariate* in the present work, referring to the patient's features.

We differentiate between prognostic and predictive biomarkers. On the one hand, *prognostic* biomarkers are clinical and biological features that affect the patient's outcome irrespective of the treatment received (Ballman, 2015). This means that prognostic features equally influence both the treated and control groups, so their information is lost when computing the treatment effect. Typical examples of prognostic biomarkers are gender, histology, or age. They are mainly used to guide aspects of clinical trial planning like patient stratification (Sechidis et al., 2018). On the other hand, a biomarker is *predictive* if it affects the treatment effect, outlining the potential benefit of the patient receiving a treatment. Therefore, predictive covariates account for the heterogeneity across the control and the treated groups. An example of a predictive biomarker is somatic mutations in oncologic diagnoses (Nalejska et al., 2014). In practice, biomarkers almost always have some degree of prognostic and predictive value but will likely be dominated by one or another. Figure 2.4 shows idealized examples of purely prognostic and purely predictive biomarkers.

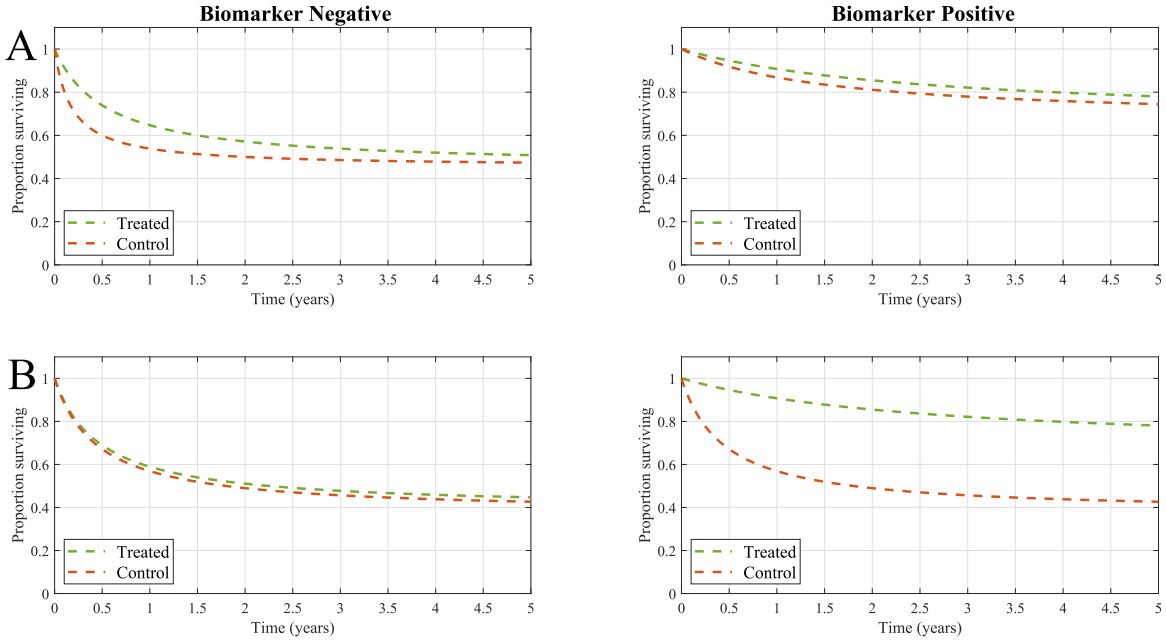



**Fig. 2.4** Experiment A shows an idealized example of a purely prognostic biomarker. The biomarker-positive patients present a better survival rate than the biomarker-negative patients, regardless of the treatment assignment (i.e., both treated and control group experiment an improvement in the survival ratio of the same magnitude). Experiment B illustrates an idealized example of a purely predictive biomarker. Here only the positive-biomarker treated group increases the surviving proportion. The patients in the control group present the same survival rate when they are biomarker-positive and biomarker-negative. Figure adapted from (Ballman, 2015).

The distinction between predictive and prognostic covariates can be ambiguous if expressed just with words. We now further clarify their difference with an example. Let us assume the additive data generation process:

$$Y = X_1 + X_2 + X_5 + W(X_3 + X_4 + X_5) + \epsilon.$$

Here, $X_1$ and $X_2$ are prognostic biomarkers – affecting the outcome $Y$ regardless of the treatment $W$ – while $X_3$ and $X_4$ are purely predictive – and only affect the treated population $W = 1$. $X_5$ is both predictive and prognostic. This is clear if we explicitly compute the difference between the POs: $Y^{(1)} - Y^{(0)} = X_3 + X_4 + X_5$, where the prognostic features cancel and only the predictive remain.

Recent ML literature has reported trends on declaring prognostic biomarkers as predictive and vice versa (Lipkovich et al., 2017; Sechidis et al., 2018). In drug discovery, mistakenly assuming a prognostic biomarker to be predictive may overestimate the benefits of a treatment for a subset of a population, resulting in financial and ethical consequences (Sechidis et al., 2018). To avoid this, a number of approaches attempt to uncover the predictive covariates by computing variable importance with some heuristics. For instance, the authors in (Svensson and Hermansson, 2021) propose several tree-based models for estimating CATE, and calculate the feature importance based on the Gini index. Our approach differs from theirs in that it can be applied to Neural Networks, which are state-of-the-art in CATE estimation. Furthermore, the theoretical formalization of our framework additionally allows finding prognostic and non-informative covariates and uncovering the covariates responsible for shifting from the general ATE to a patient-specific CATE, among other features.

We detail these features in Chapter 4, together with our proposed method. Before that, the next chapter introduces the set of tools that will allow us to interpret the CATE estimators: the saliency methods.

# Chapter 3

# Saliency methods for CATE interpretability

In the previous chapter, we discussed how individual-level feature importance has a decisive weight in the clinician's decision-making process, and we supported this with insights from human experts. Moreover, we argued that interpretability methods based on feature attribution can become a powerful tool for discovering predictive biomarkers, which are the true drivers of the treatment effect and thus of the utmost relevance for personalized medicine. We believe these two reasons are enough to base our framework on this group of interpretability methods.

To this end, this chapter focuses on *saliency methods*. Section 2.2.2 introduced saliency methods as a family of post-hoc interpretability techniques that explain the prediction of a model by highlighting the input features that are most relevant for issuing the prediction. In the CATE setting, input features correspond to patient's covariates, so that saliency methods allow identifying the covariates on which the CATE estimator relies to obtain the treatment effect.

We next introduce four well-know saliency methods, namely LIME (Ribeiro et al., 2016), DeepLIFT (Shrikumar et al., 2016), Integrated Gradients (Sundararajan et al., 2017), and SHAP (Lundberg and Lee, 2017). We do so by describing them with a unified notation, which also links to the one used in the Neyman-Rubin framework (Section 2.1.2).

**Notation** We assume a patient with covariates $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^d$ who receives a treatment $W = w$, so that we only have access to the observed outcome $Y = y$. We attempt to estimate the CATE $\tau$ with a black-box estimator $\hat{\tau}$. A saliency method will attribute an importance score $a_c(\hat{\tau}, \boldsymbol{x})$ to each covariate $c \in [d]$ contained in the vector $\boldsymbol{x}$.

## 3.1 LIME

Local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) is a pioneering work in machine learning interpretability. The authors propose an explanation technique based on a surrogate model that learns an interpretable model locally around a prediction of a black-box. The method works in a model-agnostic fashion, meaning that it can be used to explain any type of model.

In our setting, LIME computes the feature importance of a patient with covariates $\boldsymbol{x}$ by testing how giving variations of the data to the black-box CATE estimator, $\hat{\tau}$, affects its predictions. This is achieved by fitting a weighted, interpretable model $g$ in the dataset comprising the perturbed samples $\boldsymbol{z}$, which locally approximates $\hat{\tau}$. The authors argue that the explanation ensures both *interpretability* and *local fidelity*. Formally, the saliency is obtained as

$$a_1(\hat{\tau}, \boldsymbol{x}), \ldots, a_d(\hat{\tau}, \boldsymbol{x}) = \arg \min_{g \in G} \mathcal{L}\left(\hat{\tau}, g, \pi_x\right) + \Omega(g), \tag{3.1}$$

where $g$ is an interpretable model from class $G$ (e.g., linear models, decision trees), $\mathcal{L}$ is a fidelity function that measures how close is the explanation to the black-box prediction, and $\Omega$ measures the complexity of $g$ (e.g., the number of features, depth of decision tree). The proximity measure $\pi_x$ defines how close the perturbed samples $\boldsymbol{z}$ are to the instance of interest $\boldsymbol{x}$.

Specifically, the steps for computing an explanation with LIME are the following:

1. LIME first generates a new dataset made by the instance of interest $\boldsymbol{x}$ and the perturbed samples $\boldsymbol{z}$, together with their corresponding black-box predictions.
2. The new samples are weighted according to their proximity to $\boldsymbol{x}$ by $\pi_x$.
3. The weighted, interpretable model $g$ is trained on the new dataset.
4. The explanation is computed with Equation (3.1).

Figure 3.1 shows a toy example to present intuition for the method. Although model-agnostic and straightforward to use, LIME presents several drawbacks. Notably, the choice of the perturbed neighborhood can have a significant impact on the explanation, making them sometimes unreliable (Laugel et al., 2018).

## 3.2 DeepLIFT

Deep Learning Important FeaTures (DeepLIFT) (Shrikumar et al., 2016) is a gradient-based interpretability method that decomposes a prediction of a Neural Network by back-propagating

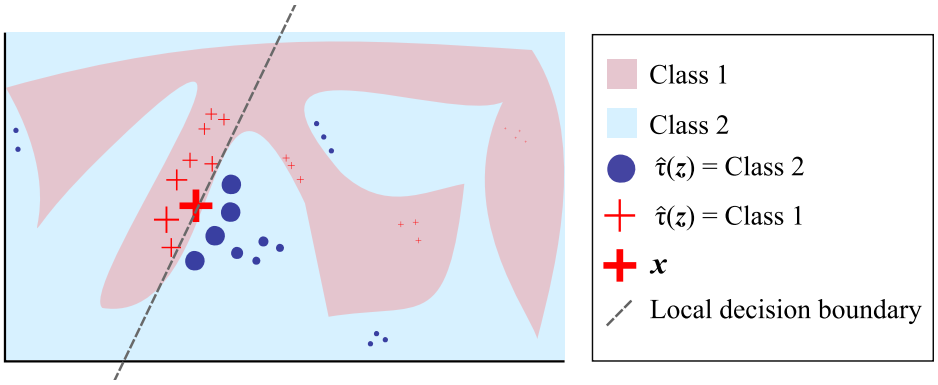



**Fig. 3.1** Toy example of a LIME explanation. $\hat{\tau}$ is a binary black-box CATE estimator with a non-linear decision boundary, represented by the pink and light blue background. The bold red cross is the instance patient $\boldsymbol{x}$ that we want to explain. LIME samples instances $\boldsymbol{z}$ by perturbing $\boldsymbol{x}$, gets their black-box predictions $\hat{\tau}(\boldsymbol{z})$, and weights them by their proximity to $\boldsymbol{x}$ – represented here by size. The gray line is the linear decision boundary for the interpretable model $g$, which explains $\hat{\tau}$ locally but not globally. Figure adapted from (Ribeiro et al., 2016).

the contributions of all the neurons to the input features. In this sense, unlike LIME, DeepLIFT is not model-agnostic. It was notably one of the first methods that introduced the concept of "baseline," framing the problem of feature importance as a difference from a reference.

We re-formulate the DeepLIFT framework presented in (Ancona et al., 2017) with our notation. Formally, DeepLIFT assigns to each neuron $i$ an attribution representing the relative impact of activating it at a patient $\boldsymbol{x}$ compared to the activation for a baseline patient $\bar{\boldsymbol{x}}$. Starting at the output layer $L$, the method first assigns a relevance $r_1^{(L)} = \hat{\tau}(\boldsymbol{x}) - \hat{\tau}(\bar{\boldsymbol{x}})$ to the output neuron[1]. Then it runs a forward pass, computing the reference values $\bar{\boldsymbol{z}}_{ji} = w_{ji}^{(l+1,l)} \bar{x}_i^{(l)}$ for all hidden units, using the baseline $\bar{\boldsymbol{x}}$ as input and recording the activations. Finally, it proceeds backwards layer by layer, assigning a relevance $r_i^{(l)}$ to each neuron $i$ of each layer $l$. It does so by applying the recursive rule:

$$r_i^{(l)} = \sum_j \frac{z_{ji} - \bar{z}_{ji}}{\sum_{i'} z_{ji'} - \sum_{i'} \bar{z}_{ji'}} r_j^{(l+1)}, \tag{3.2}$$

where $z_{ji} = w_{ji}^{(l+1,l)} x_i^{(l)}$. The feature importance for covariate $c$ is thus defined at the input layer:

$$a_c(\hat{\tau}, \boldsymbol{x}) = r_c^{(1)}. \tag{3.3}$$

The recursive algorithm in Equation (3.2) was used in the original paper, and is known as the "Rescale rule." We do not consider the "Reveal Cancel rule" (Shrikumar et al., 2017) here.

[1] In Section 2.1.3, CATE estimation was framed as a regression or binary classification problem, which consider a single neuron in the output layer. The original notation of DeepLIFT does contemplate the case of multi-output classification.

The existence of a baseline allows back-propagating the signal even when the gradients saturate, and it avoids artifacts caused by discontinuous gradients. However, the choice of the baseline is sometimes critical (Sturmfels et al., 2020). Some guidelines on defining it can be found in the original paper (Shrikumar et al., 2016) and the Distill article (Sturmfels et al., 2020)[2].

## 3.3 Integrated Gradients

Similar to DeepLIFT, Integrated Gradients (IG) (Sundararajan et al., 2017) is a gradient-based method specifically designed to interpret Neural Networks. IG frames the feature attribution problem from an axiomatic approach. The authors argue that *sensitivity* and *implementation invariance* are two fundamental axioms that should be satisfied by any interpretability methods, and they build IG under this premise.

Formally, IG defines the importance values of a feature as follows:

$$a_c(\hat{\tau}, \boldsymbol{x}) = (x_c - \bar{x}_c) \times \int_0^1 \frac{\partial \hat{\tau}\left(\bar{\boldsymbol{x}} + \alpha \times (\boldsymbol{x} - \bar{\boldsymbol{x}})\right)}{\partial x_c} d\alpha, \tag{3.4}$$

where $\bar{\boldsymbol{x}}$ is the baseline patient and $\bar{x}_c$ the corresponding baseline value for covariate $c$. Equation (3.4) shows how IG computes the importance of covariate $c$ by accumulating the gradients along a straight-line path from the baseline $\bar{\boldsymbol{x}}$ to patient $\boldsymbol{x}$. The integral can be approximated via a discrete summation. In (Sotoudeh and Thakur, 2019), the authors show that computing the Riemann sum using the trapezoidal rule can produce significantly better results than other heuristics.

IG belongs to a broader group of attribution methods called Path Methods. Each Path Method differs in how it monotonically interpolates between the two points, $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$, yielding different attribution methods for different (non-straight) paths. In fact, Path Methods are the only methods that satisfy some desirable properties, formally defined in (Friedman, 2004). The idiosyncrasy of IG is that it is the only Path Method that is additionally symmetry-preserving, meaning that it always gives the same importance to features that play the same role in the network.

[2]The article discusses different ways of choosing a baseline for Integrated Gradients, which is analogous to DeepLIFT.

## 3.4 SHAP

SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) is a model-agnostic interpretability method that explains individual predictions by means of feature importance. SHAP computes Shapley values (Datta et al., 2016; Shapley, 1953) from coalitional game theory, identifying a patient's covariates with players that act in a coalition and attributing them a score based on their contribution to the "payout", that is, to the CATE.

Shapley values measure the contribution of covariate $c$ from patient $\boldsymbol{x}$ by measuring the effect of removing the rest of covariates in all possible coalitions. Let us assume that we only have access to a subset covariates $S \subset [d]$ to estimate CATE; that is, we assume that we have removed the subset $\bar{S} = [d] \backslash S$. The appropriate way to simulate the effect of removing covariates from $\bar{S}$ is by computing the marginal expectation (Janzing et al., 2020):

$$\hat{\tau}(\boldsymbol{x}, S) = \mathbb{E}_{\boldsymbol{X} \sim \mathcal{D}} \left[ \hat{\tau} \left( \boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}} \right) \right], \tag{3.5}$$

where $\hat{\tau}(\boldsymbol{x}, S)$ denotes the predicted CATE if we only keep the subset $S$ of covariates from $\boldsymbol{x}$. The covariates from $S$ are removed by marginalization with the empirical distribution induced by the dataset $\mathcal{D}$. Since the Shapley values consider all possible coallitions, the saliency is computed as

$$a_c(\hat{\tau}, \boldsymbol{x}) = \sum_{S \subset [d] \backslash \{c\}} \frac{1}{d \begin{pmatrix} d-1 \\ |S| \end{pmatrix}} \left[ \hat{\tau}(\boldsymbol{x}, S \cup \{c\}) - \hat{\tau}(\boldsymbol{x}, S) \right]. \tag{3.6}$$

From Equation (3.6), the Shapley values for a patient $\boldsymbol{x}$ can be interpreted as the average shift created by adding covariate $x_c$ to a random subset of covariates $S$ that does not contain $x_c$.

However, computing exactly Equation (3.6) involves $2^d$ evaluations of the black-box model per patient in $\mathcal{D}$, which is intractable for big, high-dimensional datasets. In (Lundberg and Lee, 2017), the authors propose KernelSHAP, a kernel-based estimation of Shapley values inspired by local surrogate models like LIME (Section 3.1). This approach involves solving the optimization problem

$$a_1(\hat{\tau}, \boldsymbol{x}), \ldots, a_d(\hat{\tau}, \boldsymbol{x}) = \arg\min_{r_1, \ldots, r_d} \left[ \sum_{S \subset [d]} \pi_x(S) \left( \hat{\tau}(\boldsymbol{x}, S) - \hat{\tau}(\boldsymbol{x}, \emptyset) - \sum_{c \in S} r_c \right)^2 \right] \tag{3.7}$$

$$\pi_x(S) = \frac{d-1}{\begin{pmatrix} d \\ |S| \end{pmatrix} |S|(d - |S|)},$$

with $\hat{\tau}(\boldsymbol{x}, \emptyset) = \mathbb{E}_{\boldsymbol{X} \sim \mathcal{D}}[\hat{\tau}(\boldsymbol{X})]$. Note that $\hat{\tau}(\boldsymbol{x}, \emptyset)$ is equivalent to the Average Treatment Effect (ATE), and can be seen as an implicit baseline of SHAP. In this sense, SHAP quantifies how much the patient's covariates contribute to pushing the model's output from the general ATE to the patient-specific CATE. Equation (3.7) is given by recovering the Shapley values from the regression formulation of LIME in Equation (3.1), setting $\Omega(g) = 0$. The optimization problem is solved approximately by randomly drawing sets $S$ according to the density $\pi_x$ over the subsets of $[d]$.

The popularity of SHAP and of all methods that compute Shapley values is justified by their solid theoretical basis, being the unique methods that satisfy some desirable properties. In fact, Integrated Gradients (Section 3.3) also corresponds to a cost-sharing method called the Aumann-Shapley values (Aumann and Shapley, 1974). SHAP, on the other hand, is based on the Shapley-Shubik method (Shapley and Shubik, 1971), which is equivalent to averaging over multiple paths in the previously mentioned Path Methods. We refer the interested reader to (Sundararajan and Najmi, 2019), where the authors review the differences between some of the many operationalizations of the Shapley values in the context of feature importance.

This chapter described the theoretical details of four well-known saliency methods and how they relate to the CATE estimation problem. The next chapter shows how these methods can be effectively used to assist clinical decision-making under our proposed framework.

# Chapter 4

# Discovering predictive covariates with saliency methods

In this chapter, we detail our solution to interpreting CATE estimation. First, Section 4.1 formalizes the problem, and explains how saliency methods can be utilized to solve it. Next, in Section 4.2 we describe the considered CATE estimators that we aim to interpret. We implemented the models from scratch, additionally extending one of the state-of-the-art approaches from the CATE literature. Finally, in Section 4.3 we further justify the suitability of the saliency methods by highlighting how their elegant properties are desirable in the CATE setting.

## 4.1 Notation and Problem Definition

Let us consider a patient with covariates $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^d$ under the Neyman-Rubin framework (Section 2.1.2), and a black-box model $\hat{\tau}$ that estimates the corresponding CATE $\tau$. Note that the potential outcome regression surfaces can always be written as

$$\mu_w(\boldsymbol{x}) = \mu_0(\boldsymbol{x}) + w\tau(\boldsymbol{x}), \tag{4.1}$$

i.e., they share a dependence on the baseline function $\mu_0(\boldsymbol{x})$, while only $\tau(\boldsymbol{x})$ determines what is different between expected outcomes in the treated and the control group. In medicine, one distinguishes between two types of covariates that determine the two components of patient responses, namely the prognostic and the predictive covariates (Section 2.3.1). Prognostic covariates affect the outcome regardless of treatment, thus entering $\mu_0(\boldsymbol{x})$. The drivers of treatment heterogeneity are then the predictive covariates, shaping the differential responses to the treatment $W = w$ and consequently determining $\tau$. Predictive covariates are thus

information of the highest interest in precision medicine, so we argue that identifying them is the most critical goal in interpreting the CATE estimation problem.

To this end, we use the saliency methods presented in Chapter 3. By measuring the importance of the patient's covariates, saliency methods become a powerful tool for identifying the most relevant features for estimating the CATE. In an experimental setting with (1) a sufficiently powerful model and (2) enough training data, these features are, indeed, the predictive covariates.

Saliency methods can therefore disentangle the true drivers of the treatment effect for a specific patient, which is of the utmost value when designing personalized policies. Section 4.3 further validates the utility of these methods by showing how their elegant properties acquire a specific meaning in the CATE context, which can additionally support clinical decision-making.

## 4.2 Considered CATE estimators

This section describes the considered models that we aim to interpret. We select two indirect and two direct meta-learners instantiated with Neural Networks, namely TNet and SNet and XNet and DRNet.

### 4.2.1 TNet

TNet is the simplest indirect, NN-based meta-learner. It refers to the implementation of Kunzel's T-learner (Künzel et al., 2019) using feed-forward networks. Formally, TNet fits a separate network for each regression task, namely the PO surface for the treated, $\mu_1$, and the control, $\mu_0$, group. The two networks are trained using the general loss function

$$\mathcal{L}_F + \lambda \sum_{W \in \{0,1\}} \mathcal{R}\left(\Theta_{\mu_W}\right), \tag{4.2}$$

where $\mathcal{L}_F = \sum_{i=1}^{n} l\left(Y_i, \mu_{W_i}(\boldsymbol{X}_i)\right)$ is an arbitrary factual loss (e.g., cross-entropy, MSE), and $\mathcal{R}(\cdot)$ is an L2-regularizer for the network's weights, denoted as $\Theta_{\mu_W}$ with $W \in \{0, 1\}$. Once the two POs surfaces are estimated, TNet obtains the CATE by their difference:

$$\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x}). \tag{4.3}$$

In this sense, TNet fits the response variable $Y$ in Equation (2.7) by assuming that the response surfaces $\mu_w$ are group-specific, and thus depend on different conditional means $f_w\left(\cdot\right)$ and error terms $\varepsilon_w$. This implies preserving distributional differences induced by selection bias across the two groups, and allows considering different degrees of sparsity and

smoothness that vary with $W$ when regressing $Y$ against $\boldsymbol{X}$ and result in a complex CATE $\tau$ (Caron et al., 2020). TNet can be a good choice in a quasi-asymptotical setting (i.e., when the sample size tends to infinity) with a complex $\tau$ since it can estimate arbitrarily different regression surfaces, as formally derived in (Alaa and van der Schaar, 2018b).

However, this is not usually the case with real-world data. In the presence of group imbalance (e.g., in EHRs), splitting the sample involves leaving fewer observations for estimating $\mu_w$ in the smaller group. Additionally, it is a common assumption that the control and treated groups share distributional characteristics (Ballman, 2015; Curth and van der Schaar, 2021b), resulting in similar regression tasks, $\mu_0$ and $\mu_1$, and a relatively simple $\tau$. In this case, TNet underperforms other approaches that do share data between regressions.

**Algorithm 1** TNet

**Input:** $\boldsymbol{X}, Y, W$
**Output:** $\hat{\tau}$

1: $\hat{\mu}_0 = \mathrm{NN}_1\left(Y^{(0)} \sim \boldsymbol{X}^0\right)$ ▷ Estimate POs surfaces
2: $\hat{\mu}_1 = \mathrm{NN}_2\left(Y^{(1)} \sim \boldsymbol{X}^1\right)$
3: $\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x})$ ▷ Estimate CATE

### 4.2.2 XNet

XNet is a NN-based implementation of the X-learner (Künzel et al., 2019), which targets CATE directly through a multi-stage estimation procedure. It follows the three-step process detailed below.

The first step estimates the two POs surfaces, $\hat{\mu}_0$ and $\hat{\mu}_1$, by fitting an arbitrary indirect meta-learner. For simplicity, we use the TNet, described in Section 4.2.1.

In the second step, we obtain the imputed treatment effects $\tilde{D}$ by computing the difference between the observed outcomes $Y$ and the outcomes estimated with the corresponding regression surface $\hat{\mu}_w$. Formally:

$$\tilde{D}^1 = Y^{(1)} - \hat{\mu}_0\left(\boldsymbol{X}\right) \quad \text{if } W = 1 \tag{4.4}$$

$$\tilde{D}^0 = \hat{\mu}_1\left(\boldsymbol{X}\right) - Y^{(0)} \quad \text{if } W = 0, \tag{4.5}$$

for the treated and control group, respectively. Note that if $\hat{\mu}_0 = \mu_0$ and $\hat{\mu}_1 = \mu_1$ then $\tau(\boldsymbol{x}) = \mathbb{E}\left[\tilde{D}^1 \mid \boldsymbol{X} = \boldsymbol{x}\right] = \mathbb{E}\left[\tilde{D}^0 \mid \boldsymbol{X} = \boldsymbol{x}\right]$, that is, $\tilde{D}$ is an unbiased estimator of $\tau$ when $\mu_0$ and $\mu_1$ are known. Next, we estimate the group-specific CATEs, $\hat{\tau}_0$ and $\hat{\tau}_1$, in two separate non-parametric pseudo-outcome regressions, using the imputed treatment effects as the

response variable and the covariates $\boldsymbol{X}$ as regressors:

$$\tilde{D}^1 = \tau_1(\boldsymbol{X}) + \eta_1 \quad \text{if } W = 1 \tag{4.6}$$

$$\tilde{D}^0 = \tau_0(\boldsymbol{X}) + \eta_0 \quad \text{if } W = 0, \tag{4.7}$$

with $\eta_W \sim \mathcal{N}(0, \sigma^2)$. These two regressions are also instantiated with NNs.

Finally, the third step estimates the global CATE by the weighted average

$$\hat{\tau}(\boldsymbol{x}) = g(\boldsymbol{x})\hat{\tau}_0(\boldsymbol{x}) + (1 - g(\boldsymbol{x}))\hat{\tau}_1(\boldsymbol{x}). \tag{4.8}$$

As suggested by (Künzel et al., 2019), we choose the propensity score as the weighting function, i.e., $g(\boldsymbol{x}) = \pi(\boldsymbol{x})$. This requires estimating $\hat{\pi}(\boldsymbol{x})$ with a separate NN trained with a cross-entropy loss.

**Algorithm 2** XNet

**Input:** $\boldsymbol{X}, Y, W, g$
**Output:** $\hat{\tau}$

$\hat{\mu}_0 = \mathrm{NN}_1\left(Y^{(0)} \sim \boldsymbol{X}^0\right)$ ▷ Estimate response surfaces
$\hat{\mu}_1 = \mathrm{NN}_2\left(Y^{(1)} \sim \boldsymbol{X}^1\right)$

$\tilde{D}^1 = Y^{(1)} - \hat{\mu}_0\left(\boldsymbol{X}^1\right)$ ▷ Compute imputed treatment effects
$\tilde{D}^0 = \hat{\mu}_1\left(\boldsymbol{X}^0\right) - Y^{(0)}$

$\hat{\tau}_1 = \mathrm{NN}_3\left(\tilde{D}^1 \sim \boldsymbol{X}^1\right)$ ▷ Estimate group-specific CATEs
$\hat{\tau}_0 = \mathrm{NN}_4\left(\tilde{D}^0 \sim \boldsymbol{X}^0\right)$

$\hat{\tau}(\boldsymbol{x}) = g(\boldsymbol{x})\hat{\tau}_0(\boldsymbol{x}) + (1 - g(\boldsymbol{x}))\hat{\tau}_1(\boldsymbol{x})$ ▷ Average estimates

XNet solves the two problems that we identified in TNet for real-world datasets. First, it can provably adapt to structural properties of $\tau$, such as sparsity and smoothness (Künzel et al., 2019). Secondly, it is particularly effective when one treatment group is much larger than the other since it uses information from the treated group to derive estimators for the control group, and vice versa.

### 4.2.3 DRNet

The DRNet is a NN-based doubly-robust CATE estimator, introduced by Kennedy (Kennedy, 2020). It first estimates the POs surfaces and the propensity score separately, and then targets the CATE directly in a two-stage fashion.

The first step estimates the nuisance parameters $\hat{\eta}$, here the propensity score and the POs surfaces. We achieve this analogously to in XNet, using a TNet to obtain $\hat{\mu}_0$ and $\hat{\mu}_1$, and a separate NN for $\hat{\pi}$ trained with a cross-entropy loss.

The second step involves the pseudo-outcome regression. The DRNet regresses the pseuso-outcome $\tilde{Y}_{\hat{\eta}}$, based on the estimated nuisance parameters $\hat{\eta} = (\hat{\mu}_0(\boldsymbol{x}), \hat{\mu}_1(\boldsymbol{x}), \hat{\pi}(\boldsymbol{x}))$, on the covariates $\boldsymbol{X}$ by following

$$\tilde{Y}_{\hat{\eta}} = \left(\frac{W}{\hat{\pi}(\boldsymbol{X})} - \frac{(1-W)}{1-\hat{\pi}(\boldsymbol{X})}\right) Y + \left[\left(1 - \frac{W}{\hat{\pi}(\boldsymbol{X})}\right)\hat{\mu}_1(\boldsymbol{X}) - \left(1 - \frac{1-W}{1-\hat{\pi}(\boldsymbol{X})}\right)\hat{\mu}_0(\boldsymbol{X})\right]. \quad (4.9)$$

The expression in Equation (4.9) is based on the doubly-robust augmented inverse propensity weighted (AIPW) estimator (Robins and Rotnitzky, 1995). The pseudo-outcome $\tilde{Y}_{\eta}$ has the advantage of being an unbiased CATE estimator (i.e., $\mathbb{E}\left[\tilde{Y}_{\eta} \mid \boldsymbol{X} = \boldsymbol{x}\right] = \tau(\boldsymbol{x})$) if either the propensity score or the outcome surfaces are correctly predicted.

**Algorithm 3** DRNet

**Input:** $\boldsymbol{X}, Y, W$
**Output:** $\hat{\tau}$

1: $\hat{\mu}_0 = \mathrm{NN}_1\left(Y^{(0)} \sim \boldsymbol{X}^0\right)$ ▷ Estimate *nuisance* parameters
2: $\hat{\mu}_1 = \mathrm{NN}_2\left(Y^{(1)} \sim \boldsymbol{X}^1\right)$
3: $\hat{\pi} = \mathrm{NN}_3\left(W \sim \boldsymbol{X}\right)$
4: $\tilde{Y}_{\hat{\eta}} = \mathrm{AIPW\_transform}\left(Y, W, \hat{\pi}, \hat{\mu}_0, \hat{\mu}_1\right)$ ▷ AIPW transformation
5: $\hat{\tau} = \mathrm{NN}_4\left(\tilde{Y}_{\hat{\eta}} \sim \boldsymbol{X}\right)$ ▷ Estimate CATE

### 4.2.4 SNet

We use a modified version of (Curth and van der Schaar, 2021a)'s SNet. We re-implement their architecture and extend it by enforcing a balanced representation, such that the induced treated and control group distributions look similar. We show how this extension is useful for our experiments compared to standard SNet in Appendix B.1.

Let us briefly describe the SNet architecture (Curth and van der Schaar, 2021a). Formally, representation learning approaches involve jointly learning a representation $\Phi : \mathcal{X} \to \mathcal{S}$ and two regression heads $h_w : \mathcal{S} \to \mathcal{Y}$, each fit with the corresponding treated/control group data in the representation space. SNet additionally considers a third head to predict the propensity score $h_\pi : \mathcal{S} \to \mathcal{Z}$, with $\mathcal{Z} \in [0, 1]$. Note, however, that not all covariates $\boldsymbol{X}$ are useful to predict the potential outcomes and/or the propensity score. Figure 4.1b illustrates that there are at least five types of covariates depending on how they affect the treatment assignment

$W$, the observed outcome $Y$, and the two separate potential outcomes $Y^{(0)}$ and $Y^{(1)}$. SNet explicitly models these factors by learning five disentangled representations, and it estimates the POs surfaces and the propensity score by conditioning only on the relevant ones (Figure 4.1a). The model is instantiated with NNs and trained end-to-end, allowing for learning complex non-linear representations and regression heads with significant flexibility. In this sense, SNet builds on ideas from the three domains presented in Section 2.1.3: multi-task learning, representation learning, and disentangled representation learning.

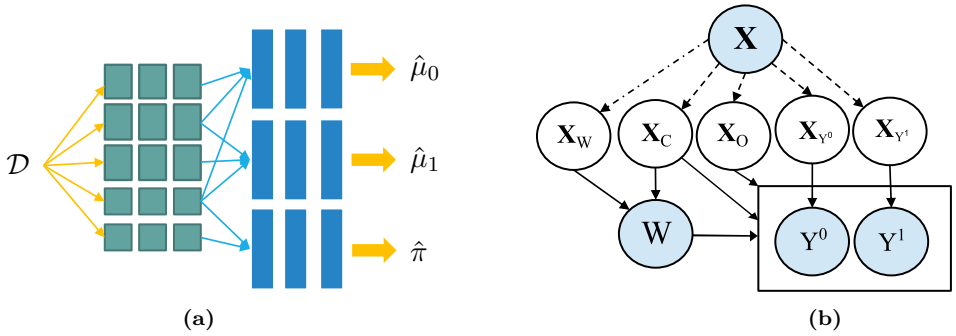



**Fig. 4.1** (a) SNet architecture (Curth and van der Schaar, 2021a). SNet is based on five disentangled representation layers – for the five assumed types of covariates – and three regression heads – for the POs and the propensity score. (b) Assumed graphical model representing the underlying factors of $\boldsymbol{X}$. $\boldsymbol{X}_O$ determines the outcome $Y$, $\boldsymbol{X}_W$ the treatment assignment $W$, $\boldsymbol{X}_C$ confounds both $Y$ and $W$, and $\boldsymbol{X}_{Y^0}$ and $\boldsymbol{X}_{Y^1}$ affect the POs $Y^0$ and $Y^1$, respectively. Selection bias is induced by $\boldsymbol{X}_W$ and $\boldsymbol{X}_C$.

These features are best seen in SNet's loss function:

$$
\begin{aligned}
J\left(h_0, h_1, h_\pi, \Phi_O, \Phi_C, \Phi_W, \Phi_{Y^0}, \Phi_{Y^1}\right) = \frac{1}{n}\Big\{ \sum_{i=1}^{n} & \underbrace{\mathcal{L}\left[h_0(\Phi_O\left(\boldsymbol{X}_i\right), \Phi_C\left(\boldsymbol{X}_i\right), \Phi_{Y^0}\left(\boldsymbol{X}_i\right)), Y_i\right]}_{\text{Factual loss } h_0} \\
& + \underbrace{\mathcal{L}\left[h_1(\Phi_O\left(\boldsymbol{X}_i\right), \Phi_C\left(\boldsymbol{X}_i\right), \Phi_{Y^1}\left(\boldsymbol{X}_i\right)), Y_i\right]}_{\text{Factual loss } h_1} \\
& + \underbrace{\text{CrossEntropy}\left[h_\pi(\Phi_W\left(\boldsymbol{X}_i\right), \Phi_C\left(\boldsymbol{X}_i\right)), W_i\right]\Big\}}_{\text{Propensity loss}} \\
& + \alpha \cdot \underbrace{\mathcal{R}_O\left(\Phi_O, \Phi_C, \Phi_W, \Phi_{Y^0}, \Phi_{Y^1}\right)}_{\text{Orthogonalization loss}} \\
& + \beta \cdot \underbrace{\mathcal{R}_I\left(\Phi_O, \Phi_{Y^0}, \Phi_{Y^1}\right)}_{\text{Imbalance loss}} \\
& + \lambda \cdot \underbrace{\mathcal{R}\left(h_0, h_1, h_\pi, \Phi_O, \Phi_C, \Phi_W, \Phi_{Y^0}, \Phi_{Y^1}\right)}_{\text{L2-regularization}},
\end{aligned}
\tag{4.10}
$$

which includes our modification to the original architecture: the imbalance loss. We now detail each of these components separately.

**Factual losses** These terms try to accomplish low-error predictions by a standard supervised learning objective (e.g., MSE, cross-entropy) using observed (factual) data. Each head is trained separately with the corresponding samples from the treated or control group, e.g., $(\boldsymbol{X}_i, W_i = 1, Y_i)$ is only used to update $h_1$. Note that the inputs for these regression networks are the outputs $\Phi_O$, $\Phi_C$ and $\Phi_{YW}$ from the representation networks. This way, the outcome estimation is conditioned only on covariates that are purely predictive of outcome $\boldsymbol{X}_O$, confounders $\boldsymbol{X}_C$, and the group-specific covariates $\boldsymbol{X}_{Y^0}$ and $\boldsymbol{X}_{Y^1}$ (see Figure 4.1b for the graphical model). Note that $\boldsymbol{X}_{Y^0}$ and $\boldsymbol{X}_{Y^1}$ are equivalent to the predictive covariates that we are interested in, so enforcing their disentanglement is a critical step in their identification.

**Propensity loss** Here we aim to minimize the error of predicting the propensity score $\pi$ by using a cross-entropy loss based on the representations that embed information about treatment assignment, $\Phi_W$ and $\Phi_C$. By minimizing this term, we ensure that learning $\Phi_W$ and $\Phi_C$ allows predicting $W$ (or, equivalently, the propensity score $\pi$).

**Orthogonalization loss** Without loss of generality, Curth et al. (Curth and van der Schaar, 2021a) assume that the observed covariates $\boldsymbol{X}$ can be decomposed into five kinds of factors $\boldsymbol{X}_W$, $\boldsymbol{X}_C$, $\boldsymbol{X}_O$, $\boldsymbol{X}_{Y^0}$, $\boldsymbol{X}_{Y^1}$ under an unknown joint distribution $\mathbb{P}(\boldsymbol{X}) = \mathbb{P}(\boldsymbol{X}_W, \boldsymbol{X}_C, \boldsymbol{X}_O, \boldsymbol{X}_{Y^0}, \boldsymbol{X}_{Y^1})$ that follows from the graphical model in Figure 4.1b. In this setting, $\boldsymbol{X}_W$ only affects treatment assignment $W$, $\boldsymbol{X}_O$ only the outcome $Y$, $\boldsymbol{X}_C$ confounds both $W$ and $Y$, and $\boldsymbol{X}_{Y^0}$ and $\boldsymbol{X}_{Y^1}$ contribute to only one potential outcome, namely the control $Y^0$ and the treated $Y^1$, respectively. Therefore, $W$ follows the distribution $\mathbb{P}(W \mid \boldsymbol{X}_W, \boldsymbol{X}_C)$, $Y$ follows $\mathbb{P}(Y \mid \boldsymbol{X}_C, \boldsymbol{X}_O)$, and $Y^0$ and $Y^1$ follow $\mathbb{P}(Y^0 \mid \boldsymbol{X}_C, \boldsymbol{X}_O, \boldsymbol{X}_{Y^0})$ and $\mathbb{P}(Y^1 \mid \boldsymbol{X}_C, \boldsymbol{X}_O, \boldsymbol{X}_{Y^1})$. The goal of the orthogonalization term is disentangling these five separate representations while ensuring their identifiability.

To this end, SNet uses an orthogonalization term inspired by Wu et al. (Wu et al., 2020) and Kuang et al. (Kuang et al., 2017) which enforces that each covariate only affects one of the five representations, thus guaranteeing identifiability[1]. The contribution of a covariate $j$ in representation $\Phi_k$ – with $k \in \{1, 2, 3, 4, 5\}$, namely the five possible representations – is approximated by $\bar{W}_{\Phi_k, j} = \sum_u \left| \boldsymbol{W}_{j,u}^{1,\Phi_k} \right|$, where $\boldsymbol{W}_{j,u}^{1,\Phi_k}$ is the first weight matrix in representation $\Phi_k$ of covariate $j$, whose weights are summed up. The orthogonalization term simply computes all possible cross-products $\bar{W}_{\Phi_k, j} \times \bar{W}_{\Phi_l, j}$ between the five different representations. This way, the regularization term $\mathcal{R}_O$ penalizes whenever a covariate enters two representations, enforcing their specialization.

[1]Since the representations and the regression functions are learned jointly, identifiability is not guaranteed otherwise.

**Imbalance loss** One of the challenges of estimating CATE from observational data is the selection bias induced by the clinician's decision-making process, which implies the treated and control populations showing different covariate distributions (details in Section 2.1.1). We extend SNet's original architecture by adding an imbalance loss term that attempts to alleviate this problem. Building on ideas from unsupervised domain adaptation (Mansour et al., 2009), several approaches propose enforcing a balanced representation by minimizing an Integral Probability Metric (IPM) (Müller, 1997; Sriperumbudur et al., 2012) distance between the two distributions. This can be understood as estimating the POs surfaces and/or the propensity score under a constraint that encourages better generalization across the treated and control groups. For two probability density functions $p$, $q$ defined over $\mathcal{R} \subseteq \mathbb{R}^d$ and for a family $\mathscr{F}$ of functions $f$, the general expression of an IPM follows

$$\mathrm{IPM}_{\mathscr{F}}(p, q) := \sup_{f \in \mathscr{F}} \left| \int_{\mathcal{R}} f(r)(p(r) - q(r))dr \right|. \tag{4.11}$$

In our experiments, we consider the family of norm-1 reproducing kernel Hilbert space (RKHS) functions, which leads to the kernel-based Maximum Mean Discrepancy (MMD) (Gretton et al., 2009) denoted as $\mathrm{MMD}(p, q)$. Another common choice in CATE estimation is the Wasserstein distance (Villani, 2009). Let us denote $\Phi^{I_C}$ and $\Phi^{I_T}$ the representations of the covariate distributions for the control and treated groups, respectively, i.e., $I_C = \{i : W_i = 0\}$ and $I_T = \{i : W_i = 1\}$. We obtain an unbiased estimation of the squared MMD distance between $\Phi^{I_C}$ and $\Phi^{I_T}$ by

$$\begin{aligned}&\mathrm{MMD}_k^2(\Phi^{I_C}, \Phi^{I_T}) = \\ &\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k\left(\boldsymbol{x}_i, \boldsymbol{x}'_j\right) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k\left(\boldsymbol{x}'_i, \boldsymbol{x}'_j\right),\end{aligned} \tag{4.12}$$

with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \sim \Phi^{I_C}, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n \sim \Phi^{I_T}$ (Appendix A.1). We choose the differentiable Gaussian RBF as kernel $k$ to ensure that the MMD is compatible with gradient-based learning:

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \exp\left(\frac{-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\sigma^2}\right). \tag{4.13}$$

We set $\sigma = 1$ in our experiments. The full imbalance term minimizes the squared MMD distance between the control and treated groups distributions for the representations of $\boldsymbol{X}_O$, $\boldsymbol{X}_{Y^0}$ and $\boldsymbol{X}_{Y^1}$:

$$\begin{aligned}\mathcal{R}_I = &\ \mathrm{MMD}^2\left(\left\{\Phi_O\left(\boldsymbol{X}_i\right)\right\}_{i:W_i=0}, \left\{\Phi_O\left(\boldsymbol{X}_i\right)\right\}_{i:W_i=1}\right) \\ &+ \mathrm{MMD}^2\left(\left\{\Phi_{Y^0}\left(\boldsymbol{X}_i\right)\right\}_{i:W_i=0}, \left\{\Phi_{Y^0}\left(\boldsymbol{X}_i\right)\right\}_{i:W_i=1}\right) \\ &+ \mathrm{MMD}^2\left(\left\{\Phi_{Y^1}\left(\boldsymbol{X}_i\right)\right\}_{i:W_i=0}, \left\{\Phi_{Y^1}\left(\boldsymbol{X}_i\right)\right\}_{i:W_i=1}\right).\end{aligned} \tag{4.14}$$

Note that we only consider the representation of covariates that affect the outcome and the POs: $\boldsymbol{X}_O$, $\boldsymbol{X}_{Y^0}$, and $\boldsymbol{X}_{Y^1}$ [2]. Artificially balancing the distributions for all covariates $\boldsymbol{X}$ – like in (Johansson et al., 2020; Shalit et al., 2017) – can counterproductively induce additional selection bias since some of the covariates ($\boldsymbol{X}_W$ and $\boldsymbol{X}_C$) embed information about treatment assignment. Therefore, the disentanglement allows us to ensure that we do not erroneously discard treatment assignment information when minimizing the loss. Curth et al. (Curth and van der Schaar, 2021a) note that these representations do not have to be invertible for causal identification – which was initially stated by (Shalit et al., 2017) – as long as they preserve all identifying conditional independence relationships, i.e., $W \perp \boldsymbol{X} \mid \Phi(\boldsymbol{X})$.

We have now described in detail the considered CATE estimators. The following section shows how we can interpret them using a set of desirable properties from the saliency methods introduced in Chapter 3.

## 4.3 Guidelines on saliency methods for interpreting CATE

Here we propose a comprehensive qualitative comparison of the saliency methods presented in Chapter 3, namely, LIME, DeepLIFT, Integrated Gradients, and SHAP. We do so by considering a set of properties that we believe are desirable in the problem of discovering predictive covariates, and we identify which interpretability methods satisfy which properties. We frame each of these properties in the CATE setting, and we show how they can support the clinical decision-making process. The properties are mainly derived from the works at (Aas et al., 2019; Ancona et al., 2017; Datta et al., 2016; Friedman, 2004; Janzing et al., 2020; Sundararajan and Najmi, 2019; Sundararajan et al., 2017).

**Baseline**. Although not a property per se, we believe that defining a baseline is desirable when measuring the importance of features. Humans assign blame to a cause by implicitly considering the absence of the cause as a baseline; that is, we perform attributions by counterfactual thinking (Roese, 1997) . In this sense, by considering feature importance as a difference-from-reference, the saliency method indeed mimics the way humans reason, making it more comprehensive and user-friendly.

DeepLIFT and Integrated Gradients allow the practitioner to select the baseline input $\bar{\boldsymbol{x}}$. For our problem at hand, we choose the average patient $\bar{\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}[\boldsymbol{X}]$. Note, however, that the average patient does not necessarily correspond to the average treatment effect (ATE). This is only guaranteed with linear models, which we already identify as interpretable. Conversely, SHAP does implicitly use as baseline the ATE (details in Section 3.4), so that it quantifies how much the patient's covariates contribute to shifting from the general ATE to the patient-specific CATE. This phenomenon is best explained with the next property.

[2]Note that $\boldsymbol{X}_O$ is independent of treatment assignment due to the collider structure at $Y$: $\boldsymbol{X}_O \rightarrow Y \leftarrow W$.

**Completeness**. Summing the attributions of all covariates gives the shift between the CATE at the patient of interest $\boldsymbol{x}$ and a baseline $b$:

$$\sum_{c=1}^{d} a_c(\hat{\tau}, \boldsymbol{x}) = \hat{\tau}(\boldsymbol{x}) - b. \tag{4.15}$$

This is particularly relevant in SHAP, where $b = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}[\hat{\tau}(\boldsymbol{X})]$. This means that the completeness property in SHAP allows measuring how much each covariate $x_c$ of patient $\boldsymbol{x}$ pushes the treatment effect from the ATE to the patient-specific CATE. This can help a clinician understand how a treatment influences an individual compared to how it affects the general population, thus discovering patient's idiosyncrasies that could help design personalized policies. We call this phenomenon ATE-Completeness. Other methods that satisfy completeness are DeepLIFT and Integrated Gradients, where $b$ is the CATE for a chosen baseline patient $\bar{\boldsymbol{x}}$; that is, $b = \hat{\tau}(\bar{\boldsymbol{x}})$ . We refer to this variation of the property as Baseline-Completeness. In fact, Integrated Gradients can be made ATE-complete by taking an expectation over the baseline with the "expected gradient", as demonstrated in (Erion et al., 2021); however, we do not explore this further in our experiments. Note that completeness implies the *sensitivity*[3] axiom in Integrated Gradients (Section 3.3). This property has also been called *Summation to Delta* (Shrikumar et al., 2016) and *Efficiency* in the context of game-theory (Lundberg and Lee, 2017; Roth, 1988).

LIME's *local accuracy* formalisation is similar to the completeness property but without considering a baseline:

$$\sum_{c=1}^{d} a_c(\hat{\tau}, \boldsymbol{x}) = \hat{\tau}(\boldsymbol{x}), \tag{4.16}$$

although it does not guarantee *sensitivity*. This means that, if the intepretable model approximates well locally the black-box predictions, LIME can reconstruct the estimated CATE $\hat{\tau}$ by adding up the contributions of all covariates.

**Dummy**. The covariates that do not affect the CATE model are given zero contribution. If $\hat{\tau}$ does not depend on covariate $c$, then $a_c(\hat{\tau}, \boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X}$. This allows detecting covariates that are non-informative for the CATE estimation. DeepLIFT, Integrated Gradients, and SHAP satisfy this property.

**Linearity**[4]. The saliency method is linear with respect to the CATE model $\hat{\tau}$. If the CATE model is expressed in terms of the estimated potential outcomes $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$, we can write:

$$a_c(\hat{\tau}, \boldsymbol{x}) = a_c\left(\hat{\mu}_1, \boldsymbol{x}\right) - a_c\left(\hat{\mu}_0, \boldsymbol{x}\right). \tag{4.17}$$

[3] *Sensitivity* implies that for every input $\boldsymbol{x}$ and baseline $\bar{\boldsymbol{x}}$ that differ in one covariate $c$ and have different CATE, $\hat{\tau}(\boldsymbol{x})$ and $\hat{\tau}(\bar{\boldsymbol{x}})$, the differing feature $c$ should be given a non-zero attribution $a_c(\hat{\tau}, \boldsymbol{x}) \neq 0$.

[4] Note that linearity cannot be applied to direct meta-learners since they target the CATE directly and do not estimate the POs surfaces.

This formulation intuitively renders the distinction between prognostic and predictive covariates. If $c$ is a prognostic covariate, one expects $a_c\left(\hat{\mu}_1, \boldsymbol{x}\right) = a_c\left(\hat{\mu}_0, \boldsymbol{x}\right) \neq 0$ so that $a_c(\hat{\tau}, \boldsymbol{x}) = 0$, which means that $c$ is not a driver of treatment effect heterogeneity. On the other hand, if $c$ is a predictive covariate, one expects $a_c\left(\hat{\mu}_1, \boldsymbol{x}\right) \neq a_c\left(\hat{\mu}_0, \boldsymbol{x}\right)$ so that $a_c(\hat{\tau}, \boldsymbol{x}) \neq 0$. This implies that $c$ is indeed relevant to explain the treatment effect heterogeneity. DeepLIFT, Integrated Gradients, and SHAP satisfy linearity.

**Symmetry-Preserving**. If $c$ and $c'$ are two symmetric[5] covariates, the saliency method assigns them the same attribution $a_c(\hat{\tau}, \boldsymbol{x}) = a_{c'}(\hat{\tau}, \boldsymbol{x})$ for all $\boldsymbol{x}$ that have identical values and identical baselines for them, i.e., $x_c = x_{c'}$ and $\bar{x}_c = \bar{x}_{c'}$. This means that if two covariates contribute equally to the estimation of CATE, they always receive the same attribution. Integrated Gradients and SHAP satisfy symmetry-preserving.

**Affine Scale Invariance (ASI)**. The attributions are invariant under a simultaneous affine transformation of the CATE $\hat{\tau}$ and the covariates $c \in [d]$. That is, for any $a$, $b$, if $\hat{\tau}_1\left(x_1, \ldots, x_d\right) = \hat{\tau}_2\left(x_1, \ldots, \left(x_c - a\right)/b, \ldots, x_d\right)$, then for all covariates $c$ we have $a_c\left(\boldsymbol{x}, \hat{\tau}_1\right) = a_c\left(\left(x_1, \ldots, a * x_c + b, \ldots, x_d\right), \hat{\tau}_2\right)$. This property implies that the zero point and the units of a covariate do not determine its importance. For instance, whether a patient's height is given in feet and inches or centimeters, the covariate receives the same attribution. Integrated Gradients and SHAP satisfy ASI.

**Implementation Invariance**. Feature attribution is only determined by the functionality of the CATE estimator, not by its specific implementation. Therefore, two functionally equivalent estimators, whose outputs are equal for all inputs despite having different implementations, should assign the same attributions. Although not directly related to the CATE setting, we believe this property is desirable for any attribution method. Integrated Gradients and SHAP are implementation invariant.

Table 4.1 summarizes the properties, outlining which saliency methods satisfy which. Note, however, that there are other factors that also influence the choice of the method. For instance, in terms of the algorithmic complexity, LIME is particularly slow since it retrains an interpretable model for each considered sample. Moreover, although an approximation of the Shapley values, the computation of SHAP is expensive for many instances since it still scales exponentially. On the other hand, Integrated Gradients only require some calls to the gradient operator of a Neural Network, which makes the method comparatively faster. Finally, DeepLIFT can be seen as an approximation of Integrated Gradients (Ancona et al., 2017), although, in practice, the differences in the execution times are not significant.

[5]Two input variables are symmetric w.r.t a function if swapping them does not change the function, i.e., if $f(x_1, x_2) = f(x_2, x_1)$ for all values of $x_1$ and $x_2$ then they are symmetric w.r.t $f$.

| | Baseline | ATE Complete | Baseline Complete | Dummy | Linear | Symmetry Preserving | ASI | Impl. Invariant |
|---|---|---|---|---|---|---|---|---|
| LIME | | | | | | | | |
| DeepLIFT | ✓ | | ✓ | ✓ | ✓ | | | |
| IG | ✓ | ✓* | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| SHAP | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 4.1** Summary of properties of the saliency methods. We do not include LIME's local accuracy since it does not imply any particularly relevant feature for CATE estimation, and is generally used just as a consistency check. * IG is not ATE-complete by definition but can be made so by computing the Expected Gradients (Erion et al., 2021).

## What's next?

We have now defined the considered CATE estimators and the saliency methods' properties. Now, how can we apply our framework to the problem of interpreting CATE?

- We can train any CATE estimator – TNet, XNet, DRNet, or SNet – on observational data and discover the predictive covariates of a separate patient $\boldsymbol{x}$ by identifying the largest attributions in the CATE saliency $\{a_c(\hat{\tau}, \boldsymbol{x})\}_{c=1}^{d}$.
- We can characterize a population by aggregating the individual feature importance of several patients to obtain *global* saliencies.
- We can detect the prognostic and the irrelevant covariates by comparing the POs surfaces' saliencies, $\{a_c(\hat{\mu}_0, \boldsymbol{x})\}_{c=1}^{d}$ and $\{a_c(\hat{\mu}_1, \boldsymbol{x})\}_{c=1}^{d}$, in indirect CATE estimators (TNet and SNet).
- We can implicitly estimate the ATE for a population and explain how it differs from the CATE for a specific patient $\hat{\tau}(\boldsymbol{x})$ based on the feature attributions $\{a_c(\hat{\tau}, \boldsymbol{x})\}_{c=1}^{d}$, thanks to the ATE-Completeness property in SHAP.
- We can attribute the difference between the CATE for a particular patient $\hat{\tau}(\boldsymbol{x})$ and the CATE for the average patient $\hat{\tau}(\bar{\boldsymbol{x}})$ to specific features by applying the Baseline-Completeness property of IG and DeepLIFT.
- We can guarantee the robustness of the estimated attributions by their solid theoretical formalization in the form of properties like Implementation Invariance, Symmetry Preserving, ASI, or Dummy.

We extensively investigate these features in the next chapter.

# Chapter 5

# Experimental Validation

This chapter provides an extensive evaluation of the framework presented in Chapter 4. First, in Section 5.2 we further compare the saliency methods, now quantitatively, and we illustrate their practicality by interpreting two examples of CATE estimation problems, one consisting on real-world data. Next, in Section 5.3 we fix a saliency method and compare the models described in Section 4.2 based on their ability to disentangle the predictive features. We do so in a range of synthetic datasets that are inspired by the drug development setting.

## 5.1 Implementation Details

We implement the CATE estimators using similar specifications as in (Curth and van der Schaar, 2021a; Shalit et al., 2017). We use fully-connected layers with exponential linear units (ELU) as non-linear activation functions. In TNet, we set 2 layers with 200 neurons and 100 neurons for each regression head and an output layer with a single neuron, which includes a sigmoidal function if the output is binary. The first stage for DRNet and XNet estimates the POs surfaces using a TNet with that same characteristics. We also set 2 layers with 200 and 100 neurons for the second stage, both for estimating the propensity score – with an additional sigmoid activation – and for regressing the specific pseudo-outcomes. For SNet, we consider two scenarios: when there is selection bias – so we include representations $\Phi_C$ and $\Phi_W$ and the propensity head – and when there is not, so we exclude them. For the first scenario, we set 1 layer for each shared representation, with 50 neurons for $\Phi_O$, $\Phi_{Y^0}$ and $\Phi_{Y^1}$, and 100 neurons for $\Phi_W$ and $\Phi_C$. For the second scenario, we also set 1 layer for each shared representation, now with 50 neurons for $\Phi_{Y^0}$ and $\Phi_{Y^1}$, and 100 neurons for $\Phi_O$. The regression heads in both cases use 1 layer with 100 neurons. We consider selection bias in simulations with the Twins dataset (Section 5.2.4) and with DGP7 in Section 5.3.2.

These specifications ensure that each estimated function $(\hat{\mu}_w, \hat{\pi}, \hat{\tau})$ has access to the same amount of layers and neurons, and therefore that each architecture can represent equally complex functions. We train the models for (max.) 10.000 epochs with Adam (Kingma and Welling, 2013), learning rate lr=0.0001, mini-batches of size 300, early stopping based on a 30% validation, and a patience of 10 epochs. We set an L2-penalty $\lambda = 0.01/100$ and an orthogonalization penalty $\alpha = 0.1$ for all models. We tune the imbalance penalty $\beta$ and set it to $\beta = 1$ (Appendix B.1). Note that the imbalance loss term is only considered in the selection bias scenario. All models are implemented from scratch in PyTorch (Paszke et al., 2019).

We now specify the design choices for the saliency methods. In LIME, we use a linear regression as interpretable model $g$. For the proximity function $\pi_x$ we choose the euclidean distance between the original and the perturbed samples and we pass it through an exponential kernel with $\sigma = 2$. We measure LIME's complexity term $\Omega$ with the number of non-zero features, and we train $g$ with 1,000 samples obtained by perturbing features of the instance of interest, selected uniformly at random. Besides, DeepLIFT attributions are computed with the Rescale rule, and we approximate the attributions for Integrated Gradients using 1,000 steps of a Riemann sum based on the trapezoidal rule. Finally, KernelSHAP is approximated with 1,000 steps. We choose the number of steps/samples for LIME, Integrated Gradients and SHAP based on consistency checks (Appendix B.2).

## 5.2 On the suitability of saliency methods for interpreting CATE

We first compare the saliency methods by assessing the fidelity of their importance scores in Section 5.2.2. To make sure that errors and virtues are attributed to the saliency method and not to the model, we need (1) a sufficiently powerful model, (2) enough training data, and (3) a well-defined data generation process (DGP). We choose SNet as the CATE estimator to fulfill the first requirement, which we demonstrate in Section 5.3.2 to be the most accurate of the considered models on average for simple outcome surfaces. Requirements (2) and (3) are satisfied by designing our own DGP, so that we can query it unlimitedly and have access to its ground truth saliency.

Next, Sections 5.2.3 and 5.2.4 provide the reader with guidelines on how to use the saliency methods based on the properties described in Section 4.3. We illustrate this with two use-cases: a semi-synthetic dataset – where we simulate the (potential) outcomes based on real covariates – and a real-world one.

### 5.2.1 Datasets and Metrics

An inherent property of causal inference datasets is that counterfactual outcomes are never observed, which implies that we do not have access to the true CATE to evaluate a proposed estimator. The standard solution is to use synthetic data where the POs for all possible treatments are available. We do so in a fully-synthetic dataset to compare the saliency methods (Section 5.2.2) and in our modified version of the IHDP dataset for our first use-case (Section 5.2.3) . An exception to this is the Twins dataset, where we do have access to the two POs and therefore to the ground truth treatment effect. We use Twins for our second use-case (Section 5.2.4).

**Fully-synthetic dataset** We compare LIME, DeepLIFT, Integrated Gradients, and SHAP quantitatively in Section 5.2.2. We do so in the fairly simple DGP:

$$Y = 5X_1 + 10X_2 + 5X_3 + 3X_4 + W\left(X_5 + \frac{1}{2}X_6\right) + \epsilon,$$

with $\epsilon \sim N(0, 0.1)$. We consider uncorrelated covariates sampled from a multivariate Gaussian distribution, and no irrelevant features (i.e., all covariates are either predictive or prognostic). We set a small number of features so that each mistake has a more significant impact on average.

**Modified IHDP Dataset** The Infant Health and Development Program (IHDP) dataset has frequently been used to evaluate ML approaches to causal effect estimation (Hill, 2011; Johansson et al., 2020). The original data comes from a randomized study of the impact of early educational and follow-up interventions to enhance the cognitive, behavioral, and health status of premature infants (Brooks-Gunn et al., 1992). Each observation represents a single child in terms of 25 covariates (6 continuous and 19 binary) of their birth and their mother. We use the data prepared by Hill (Hill, 2011), where she removed a subset of the original population, namely all treated children with non-white mothers, leaving a complete dataset with 747 subjects. The data used by Hill presents both confounding and lack of overlap, and none of the covariates has a purely predictive effect, which is not ideal for evaluating the saliency methods. Therefore, we keep the covariates information but assume an RCT instead, where all individuals have the same probability of being in the treated and the control group (i.e., $\pi(\boldsymbol{x}) = 0.5, \forall \boldsymbol{x} \in \mathcal{X}$), and assign the treatment as $W \sim \text{Bern}\,(0.5)$. We then synthesize the outcome according to an additive DGP on the covariate set. In our experiments, we consider the arbitrary response surface

$$Y = 2X_1 + 3X_2 - 5X_3 + 3X_5 + 5\sum_{j=7}^{9} X_j - 5X_{13}^2 - 3X_{15} - 2X_{16} + W\left(X_{11} + \frac{1}{2}X_{12} + X_{18}^2 + \frac{1}{4}X_{25}\right) + \epsilon$$

with $\epsilon \sim N(0, 0.1)$, where the difference between predictive and prognostic information is clear (details in Appendix C.3). $Y$ can be understood as a continuous metric representing the cognitive test score (e.g., Intelligence Quotient). Note that here we not only have access to the ground truth CATE but also to the true saliencies, that is, which covariates are prognostic, predictive, or irrelevant. The full description is given in Table C.1 in Appendix C.1.

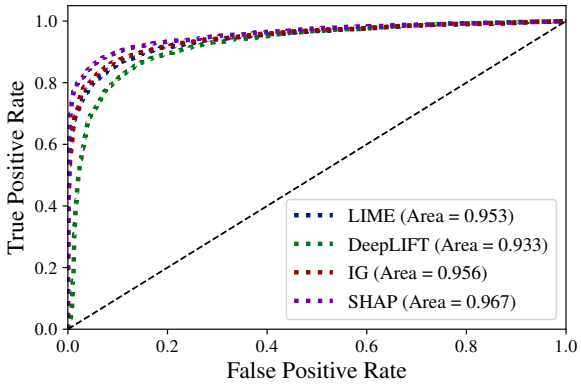
**Twins Dataset** The Twins dataset (Louizos et al., 2017) is derived from all births in the USA between 1989-1991 (Almond et al., 2005). Among these births, the authors of the dataset focused on the twins. They defined the treatment $W = 1$ as being the heavier twin and thus $W = 0$ as being the lighter twin, and the 1-year mortality as the binary outcome. For each twin pair, 30 covariates are obtained relating to their parents, the pregnancy, and the birth. In our setting, we only choose twins that are the same sex and that weight less than 2kg, while ensuring no missing covariates. The final cohort comprises 11,400 pairs of twins. This way, for each twin pair, we observe both the case $W = 0$ (lighter twin) and $W = 1$ (heavier twin); therefore, we have access to the ground truth of the CATE. In order to simulate an observational study, we selectively observe one of the two twins, using the covariate information and inducing selection bias. The process follows: $W \mid \boldsymbol{X} \sim \text{Bern}\left(\sigma\left(\boldsymbol{X}\mathbf{w} + n\right)\right)$ where $\mathbf{w} \sim \mathcal{U}\left((-0.1, 0.1)^{39\times 1}\right)$ and $n \sim \mathcal{N}(0, 0.1)$. We use 39 covariates since we one-hot-encode three of the original features. The full description is given in Table C.2 in Appendix C.2.

**Metrics** We evaluate the fidelity of the feature attributions with the Area Under the Receiver Operating curve (AUROC). The ROC curve allows evaluating the discrimination power of a binary classifier as its threshold is varied, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. We extend the ROC curve to our setup by considering each feature attribution as a binary prediction and comparing it to a binarized ground truth feature importance (i.e., 0: prognostic/irrelevant feature, 1: predictive feature). Therefore, given a matrix with a set of patients and their feature attributions, we construct the ROC curve by looking at all the samples together – all features of all patients – and computing the TPRs and FPRs for various thresholds (micro-averaging). The AUROC is then computed approximating the integral of the area under the ROC curve with the trapezoidal rule. The higher the AUROC, the better. Note that to measure the AUROC we need the ground truth saliencies, which is only possible in synthetic datasets.

### 5.2.2 Comparing saliency methods quantitatively

This section compares LIME, DeepLIFT, Integrated Gradients, and SHAP based on the fidelity of their feature attributions – which should give the predictive features the largest importance. We guarantee that SNet is well trained by using a reasonable amount of 5,000 samples from the proposed fully-synthetic dataset. This is to ensure that we indeed measure the performance of the saliency method and not the estimator's. We evaluate the saliency methods using the AUROC metric in an independent test set of 1,000 synthetic samples from the same dataset.

Figure 5.1 shows the results. All saliency methods perform very well in terms of the AUROC score, which is not surprising given the simplicity of the DGP. We can conclude that, if we use a sufficiently powerful CATE estimator and enough training data to fit it, the relevant features discovered by the saliency methods are usually the predictive features.




**(a)** ROC curves

| | **LIME** | **DeepLIFT** | **IG** | **SHAP** |
|---|---|---|---|---|
| **AUROC** | 0.953 | 0.933 | 0.956 | 0.967 |

**(b)** AUROC scores

**Fig. 5.1** Quantitative comparison of saliency methods. The closer the ROC curve to the top-left corner and the higher the AUROC, the better.

### 5.2.3 Infant Health and Development Program (IHDP)

We train SNet in a subset of the IHDP data, using the original covariates as inputs and our simulated responses as (observed) outcomes. The subset comprises 90% of the dataset, and we ensure a balanced representativeness of individuals from both treatment groups. We then compute the saliencies of a randomly selected patient from the remaining 10% of the data. To make this example as natural and intuitive as possible, we keep the names from the covariates

of the original IHDP experiment (see Table C.1 in Appendix C.1). Note, however, that the simulated response surface is arbitrary and does not aspire to mimic a realistic setting.

Figure 5.2 shows the attributions obtained with the four saliency methods. DeepLIFT, IG, and SHAP successfully disentangle the predictive features: whether the mother went to high school ($X_{11}$), whether she went to university ($X_{12}$), whether she received prenatal care ($X_{18}$), and whether the data instance was collected by the University of Washington ($X_{25}$). A practitioner could then conclude, for example, that prenatal care is a driver of treatment-effect heterogeneity. This means that children whose mothers received prenatal care increased their cognitive test scores after the IHDP interventions compared to children that did not join the program, despite (their mothers) receiving prenatal care. Therefore, if a mother received prenatal care, it would be beneficial for her premature children to join a program similar to the IHDP to increase their cognitive test scores.

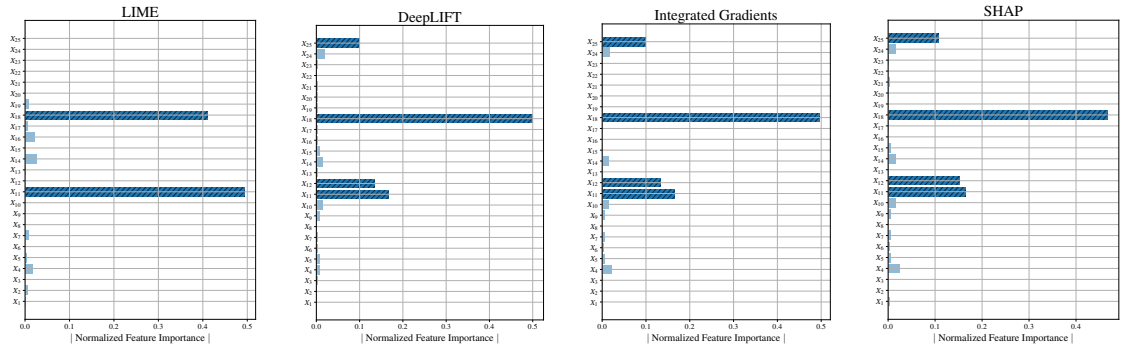



**Fig. 5.2** Absolute values of the normalized feature attribution obtained with LIME, DeepLIFT, IG and SHAP for the considered child. Predictive features are highlighted in darker blue. Normalization is performed by min-max scaling, and it allows comparison across methods.

LIME, however, can only discover two of the predictive features. This phenomenon suggests that, although LIME performed well when interpreting a well-trained estimator (Section 5.2.2), it is a less robust saliency method when the data is scarcer – which is the case of the IHDP dataset with 747 individuals. The lack of fidelity of LIME can be further verified by testing its local accuracy $\sum_{c=1}^{d} a_c(\hat{\tau}, \boldsymbol{x}) = \hat{\tau}(\boldsymbol{x})$. Here, the sum of the raw, unnormalized attributions results in $\sum_{c=1}^{25} a_c(\hat{\tau}, \boldsymbol{x}) = 0.116$, which is far from SNet's prediction $\hat{\tau}(\boldsymbol{x}) = 0.139$.

We illustrate the utility of the linearity property with Figure 5.3 using Integrated Gradients. Here we do not normalize or compute the absolute value of the feature attributions. Conversely, we visualize how each covariate influences the estimator's prediction for the child compared to the prediction for the baseline child $\bar{\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}[\boldsymbol{X}]$. By comparing the saliencies for $\hat{\mu}_0$ and $\hat{\mu}_1$, we can identify the prognostic and irrelevant covariates for child $\boldsymbol{x}$. For instance, the birth weight ($X_1$), the number of weeks born preterm ($X_2$), and whether the mother

smoked during pregnancy ($X_{15}$), among others, are purely prognostic since they affect equally both POs surfaces, which implies that they influence the child's cognitive score regardless of him/her joining the program. For the considered child, the mother did smoke (i.e., $x_{15} = 1$), which decreases the cognitive test scores compared to the average child – whose mother does not smoke (i.e., $\bar{x}_{15} = 0$). An example of an irrelevant feature is the mother's age ($X_6$), which is given a close-to-zero attribution for both POs surfaces. Finally, predictive features only affect $\hat{\mu}_1$ and thus remain in the CATE $\hat{\tau}$'s saliency. Looking at the effect's sign, the considered child presents values that are harmful for the test scores in all predictive covariates. Specifically, the mother did not graduate from high school ($x_{11} = 0$), did not graduate from university ($x_{12} = 0$), did not receive any prenatal care ($x_{18} = 0$), and the data instance was not collected by the University of Washington ($x_{25} = 0$). These values decrease the potential gain in cognitive test scores compared to the baseline patient, who does fulfill these features (e.g., the mother graduated from university).

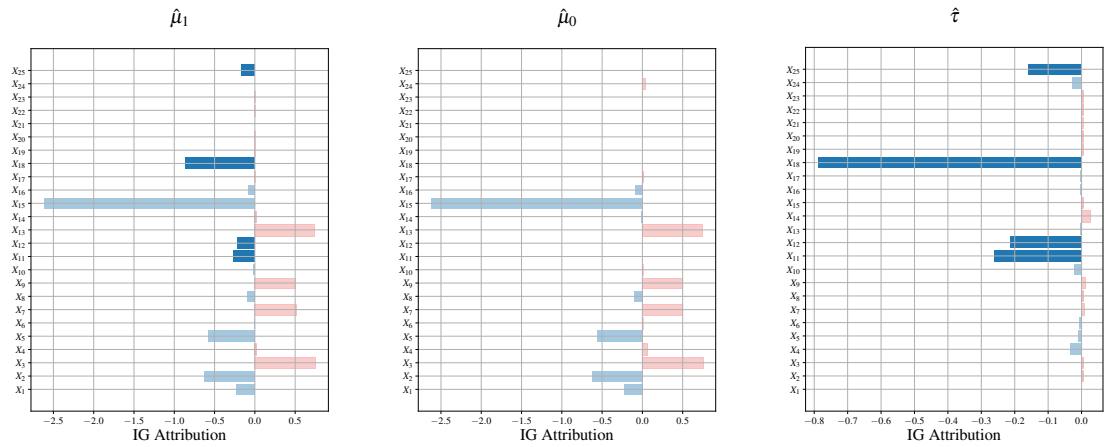



**Fig. 5.3** Feature attributions obtained with Integrated Gradients for the estimated POs surfaces and CATE. The axis is centered at 0, and the bars measure the influence of each covariate on $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\tau}$ – for a child $\boldsymbol{x}$ – compared to the estimations for a baseline $\bar{\boldsymbol{x}}$, which we choose to be the average child. Blue means a negative impact and red a positive impact. Note the change in scale in $\hat{\tau}$'s plot.

Finally, Figure 5.4 shows a force plot with the attributions obtained with SHAP. Here we can see how the specific covariates' values shift the output of SNet from SHAP's baseline – the Average Treatment Effect (ATE) – to the child-specific CATE $\hat{\tau}(\boldsymbol{x})$. We verify some of our previous conclusions. The estimated CATE for the considered child is lower than the estimated ATE; that is, the increase in cognitive test scores is lower than the average population effect. This is primarily due to the mother not having received prenatal care. The lack of educational background also severely decreases the effect of participating in the program, followed by the fact that the data does not come from the University of Washington. Note that if any of these features was actionable, it would allow designing personalized

policies for the child to increase the cognitive test scores compared to the average treatment effect.

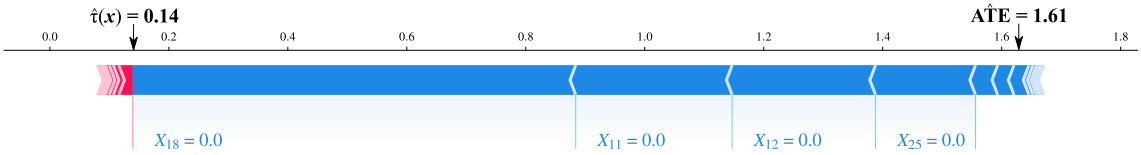



**Fig. 5.4** Force plot illustrating how the feature attributions shift the SNet's prediction from $\hat{\text{ATE}}$ to the CATE $\hat{\tau}(\boldsymbol{x})$. Blue means a negative impact and red a positive impact. We indicate the value of the relevant covariates for $\boldsymbol{x}$. Here all predictive covariates harm the treatment effect compared to the ATE. Note that they are all binary covariates.

### 5.2.4 Twins

There are various reasons why the problem of interpreting CATE is challenging for the Twins dataset. First, we do not know the underlying expression of the response surface; therefore, we cannot evaluate if the obtained saliencies are correct. Second, it is highly probable that no covariate will be purely predictive or purely prognostic, making the analysis of the saliencies harder. Third, due to the low mortality, the signal for CATE estimation is weak and noisy. Consequently, here we tackle the problem from the clinicians' perspective, without access to the ground truth. Recall that the output is the 1-year mortality. Here we consider the raw probability from the sigmoid instead of a binary output since it allows a clearer interpretation. The estimated CATE is thus a continuous value $\hat{\tau}(\boldsymbol{x}) \in [-1, 1]$, representing the increase/decrease in the probability of not surviving the first year under the treatment of "being heavier at birth." We train SNet in a balanced subset with 90% of the data and compute the saliencies for a randomly drawn patient from the remaining 10%.

Figure 5.5 shows the saliencies for the estimated POs regressions and the CATE, obtained with DeepLIFT. Positive attributions imply an increase in the probability of not surviving, and the opposite for negative attributions, compared to the (baseline) average child $\bar{\boldsymbol{x}}$. We identify some degree of *prognosticness* in covariates $X_1$ (mother's age), $X_5$ (weight gain during pregnancy, in pounds), and $X_6$ (weeks of gestation), having similar importance in both POs surfaces (i.e., in both twins). The considered child (i.e., twin pair) presents a higher weight gain but a lower number of gestation weeks than the baseline, specifically $x_5 = 28$, $x_6 = 28$ and $\bar{x}_5 = 25.4$, $\bar{x}_6 = 31.5$. Having a higher weight is found to decrease the mortality risk by DeepLIFT, while a shorter gestation period seems to be harmful to the child, increasing the risk. Furthermore, the most relevant covariates for the CATE – and thus the covariates that encode some predictive information – are $X_1$, $X_5$, $X_6$, $X_2$ (month prenatal care began), and $X_8$ (the number of prenatal visits). We associate the negative feature attribution given

to the (low) number of gestation weeks ($X_6$) – which decreases the 1-year mortality at the CATE level – to the fact that it is more harmful to the lighter twin than to the heavier, and thus the difference of saliencies is negative. The same happens with the weight gain ($X_5$) but with the opposite effect, meaning that having a higher weight gain than the average child benefits less the heavier twin than the lighter twin. Very similar conclusions can be drawn from the rest of the covariates. For instance, the prenatal care ($X_2$) for the considered child began two months later than for the average child, which results in a positive attribution at the POs level – increasing the probability of not surviving – but a negative attribution at the CATE level, since it affects the lighter twin more. Note that if it were not for the linearity property, interpreting the CATE $\hat{\tau}$'s saliency would have been considerably ambiguous.

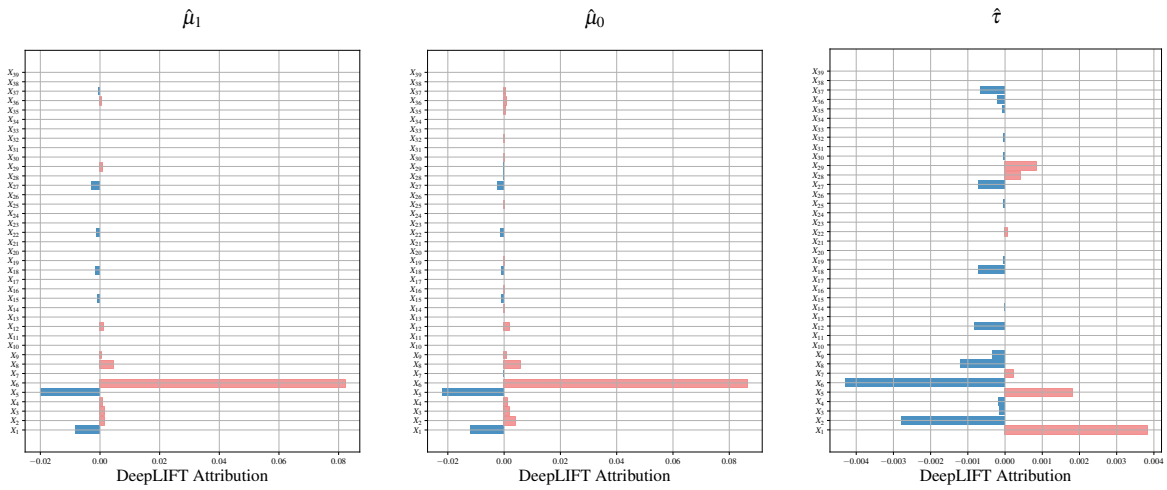



**Fig. 5.5** Feature attributions obtained with DeepLIFT for the estimated POs surfaces and CATE. The axis is centered at 0, and the bars measure the influence of each covariate on $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\tau}$ – for a child $\boldsymbol{x}$ – compared to the baseline child $\bar{\boldsymbol{x}}$. Note the change in scale in $\hat{\tau}$'s plot.

Finally, Figure 5.6 illustrates how the feature attributions shift the CATE estimation from the baseline to the child of interest. We consider SHAP and IG, which use the ATE and the CATE for the average child as baseline, respectively. IG draws similar results to DeepLIFT, giving the largest attribution to the weeks of gestation ($X_6$), the mother's age ($X_1$), and the prenatal care time ($X_2$). Interestingly, SHAP does not give a significant attribution to the weeks of gestation and ranks as predictive covariates $X_{27}$ (the mother having a medical risk factor that is not considered in other covariates) and $X_{29}$ (having intermediate adequacy of care). This is due to the methods using different baselines. For both saliency methods, the specific covariates help to decrease the mortality risk from the baseline to the considered child.

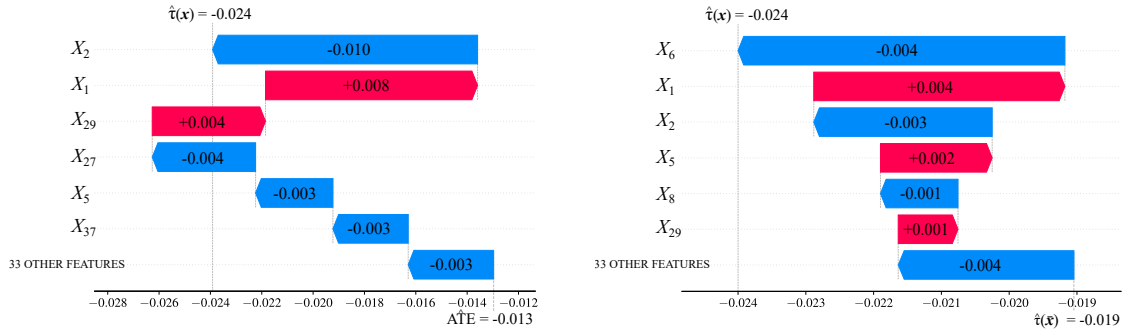



**Fig. 5.6** Waterfall plot illustrating the shift from the child's CATE $\hat{\tau}(\boldsymbol{x})$ to the baseline, obtained with (a) SHAP and (b) Integrated Gradients. We show the values for the most relevant covariates, the estimated CATE, and the corresponding baseline.

## 5.3 Characterizing CATE estimators with SHAP

We now change the axis of comparison. We fix SHAP as saliency method and experiment with various CATE estimators in a set of synthetic datasets. We choose SHAP because it showed the best performance in terms of the AUROC score in Section 5.2.2. We experiment with four estimators: TNet, SNet, XNet, and DRNet (see Section 2.1.3 for details). Specifically, we study how reliable they are in disentangling the predictive features based on the saliencies obtained with SHAP. Note that synthetic data is required in these experiments since we need to know the true response surfaces to evaluate if the saliencies are correct.

### 5.3.1 Datasets and Metrics

**Synthetic Datasets** We design our DGPs inspired by the drug development setting, particularly by the works in (Sechidis et al., 2018). The number of covariates is fixed to $d = 20$, which is a sensible, common number in these settings (Alemayehu et al., 2018; Lipkovich et al., 2017; Loh et al., 2019; Zhang et al., 2018). We include many prognostic covariates with a stronger magnitude than the predictive ones, as well as non-informative covariates that are neither prognostic nor predictive. Again, this reflects the reality in clinical trials. We assume fairly challenging scenarios, with fully-separated predictive and prognostic information and correlated covariates sampled from a multivariate Gaussian. We provide the details of the DGPs in Appendix C.3.

We show the considered DGPs in Table 5.1. We first assume a Randomized Control Trial (RCT), with neither covariates that are predictive of treatment assignment nor confounders, and hence no selection bias to account for (i.e., $\mathbb{P}(W \mid \boldsymbol{X}) = \mathbb{P}(W)$). The DGPs contemplate scenarios with an enhanced treatment effect for a sub-population, many prognostic covariates

(DGP2), strong prognostic effects (DGP3), and non-linearities and interactions (DGP4, DGP5). We additionally experiment with increasing the overall treatment effect in the population $c$ (DGP6) and with including $n_c$ confounders that affect treatment assignment (DGP7) – and thus induce selection bias. We experiment with different training set sizes. We simulate 10 runs for each DGP, and we report mean and standard deviation of the *global* feature importance given by SHAP for 100 generated test observations. We compute the global feature importance by naively aggregating the absolute values of the individual attributions and computing the average.

| | |
|---|---|
| DGP1 | $Y = 2 + 3\sum_{j=1}^{5} X_j + W\left[\mathbb{I}\left(X_7 > 0 \cap X_8 > 0\right) + \mathbb{I}\left(X_{20} > 0\right)\right] + \epsilon$ |
| DGP2 | $Y = 2 + 10\sum_{j=1}^{12} X_j + W\left[\mathbb{I}\left(X_{13} > 0 \cap X_{14} > 0\right) + \mathbb{I}\left(X_{20} > 0\right)\right] + \epsilon$ |
| DGP3 | $Y = 2 + 25\sum_{j=1}^{6} X_j + W\left[\mathbb{I}\left(X_7 > 0 \cap X_8 > 0\right) + \mathbb{I}\left(X_{20} > 0\right)\right] + \epsilon$ |
| DGP4 | $Y = 2 + 3\left(X_2 + X_3 + X_4\left(X_5 + X_6 X_7\right)\right) + W\left[\mathbb{I}\left(X_8 > 0 \cap X_9 > 0\right) + X_{10} + X_{11}^2\right] + \epsilon$ |
| DGP5 | $Y = 2 + 5\left(X_1 + X_2 X_3 + X_4^2 + X_5 X_6 X_7\right) + W\left[\mathbb{I}\left(X_8 > 0 \cap X_9 > 0 \cap X_{10} > 0\right) + X_{11} X_{12}\right] + \epsilon$ |
| DGP6 | $Y = c + 5\sum_{j=1}^{5} X_j + W\left[\cos(X_6) + \mathbb{I}\left(X_7 > 0 \cap X_8 > 0\right) + X_{10}^2\right] + \epsilon$ |
| DGP7 | $Y = 2 + 10\sum_{j=1}^{5+n_c} X_j + W\left[\mathbb{I}\left(X_{16} > 0\right) + \mathbb{I}\left(X_{17} > 0 \cap X_{18} > 0 \cap X_{19} > 0\right) + \mathbb{I}\left(X_{20} > 0\right)\right] + \epsilon$ |

**Table 5.1** Considered data generation process (DGP) models. Appendix C.3 gives the details of the simulations. All DGPs enhance the treatment effect for a sub-population due to the thresholding indicator function in the predictive component. For all DGPs $\epsilon \sim N(0, 0.1)$.

**Metrics** We track the expected *Precision in Estimation of Heterogeneous Effect* (PEHE) (Hill, 2011):

$$\epsilon_{PEHE} = \int_{\mathcal{X}} \left(\hat{\tau}(\boldsymbol{x}) - \tau(\boldsymbol{x})\right)^2 p(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\tau}(\boldsymbol{x}_i) - \tau(\boldsymbol{x}_i)\right)^2, \tag{5.1}$$

equivalent to the Mean Squared Error (MSE). We evaluate the ability to disentangle predictive features by a heuristic true positive rate (TPR)[1] (Sechidis et al., 2018):

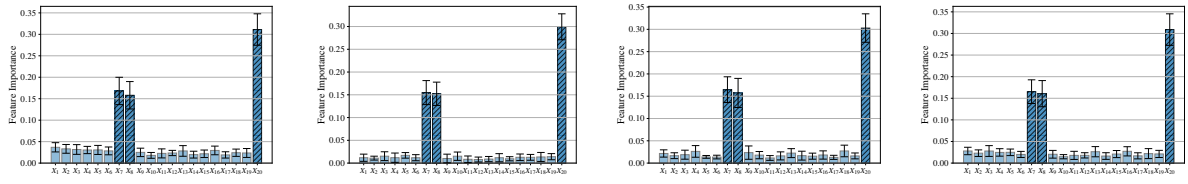
$$\text{TPR} = \frac{\left|\boldsymbol{X}_{\text{Pred}} \cap \widehat{\boldsymbol{X}}_{\text{Pred}}^{\bar{K}}\right|}{K}, \tag{5.2}$$

where $\boldsymbol{X}_{\text{Pred}}$ is the set of predictive covariates, with size $K = |\boldsymbol{X}_{\text{Pred}}|$, and $\widehat{\boldsymbol{X}}_{\text{Pred}}^{K}$ is the set of the top-$K$ covariates returned by the estimator according to their importance ranking. The TPR thus represents the fraction of predictive biomarkers correctly ranked in the top-$K$ positions, being $K$ the number of predictive features. Both metrics are computed for the 100 generated test observations. The lower the $\epsilon_{PEHE}$ and the higher the TPR, the better.

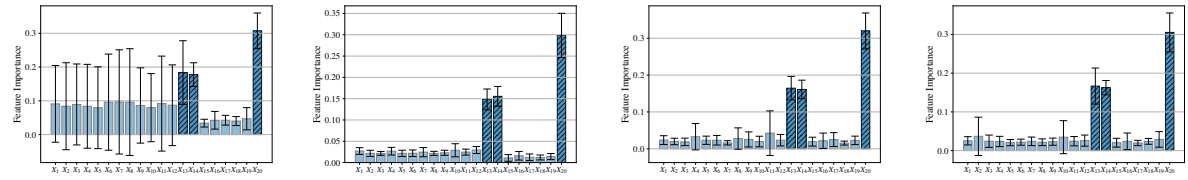
[1] This heuristic TPR is equivalent to the average R-precision in document retrieval (Manning et al., 2008).

### 5.3.2 Simulated drug development settings

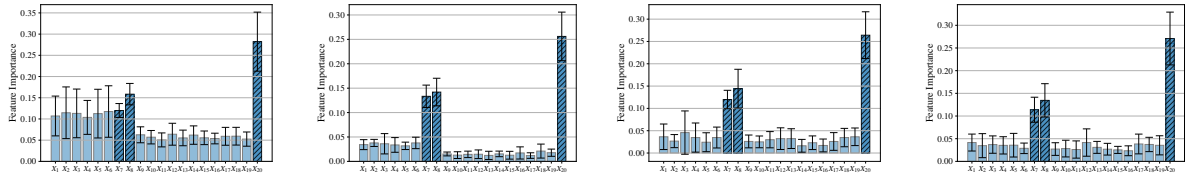
**Preliminary findings** We base our preliminary findings on the (simpler) linear synthetic datasets: DGP1, DGP2, and DGP3 (Figure 5.7). This way, we study how increasing the power and/or the number of prognostic covariates affect the ability to disentangle the predictive features. Figure 5.7 shows the attributed feature importance from each of the CATE estimators when trained with 2,500 samples. We observe that SNet, DRNet, and XNet successfully discover the predictive covariates in the three scenarios. TNet, however, appears to show bias towards prognostic biomarkers, often ranking them as predictive. This is particularly notable when we substantially increase the prognostic effect in DGP3 (Figure 5.7c). Note that non-informative covariates are never ranked as predictive by any estimator.




**(a)** DGP1: $Y = 2 + 3\sum_{j=1}^{5} X_j + W\left[\mathbb{I}\left(X_7 > 0 \cap X_8 > 0\right) + \mathbb{I}\left(X_{20} > 0\right)\right] + \epsilon$




**(b)** DGP2: $Y = 2 + 10\sum_{j=1}^{12} X_j + W\left[\mathbb{I}\left(X_{13} > 0 \cap X_{14} > 0\right) + \mathbb{I}\left(X_{20} > 0\right)\right] + \epsilon$




**(c)** DGP3: $Y = 2 + 25\sum_{j=1}^{6} X_j + W\left[\mathbb{I}\left(X_7 > 0 \cap X_8 > 0\right) + \mathbb{I}\left(X_{20} > 0\right)\right] + \epsilon$

**Fig. 5.7** Global feature importance given by SHAP for DGP1, DGP2, and DGP3. TNet, SNet, DRNet, and XNet are trained with 2,500 samples, and the feature importance is given for a separate test set of 100 samples. We report the mean and standard deviation for 10 simulations of each DGP. We highlight the true predictive features.

Figure 5.8 (DGP2) and Figure 5.9 (DGP3) conclude that SNet is the most sample-efficient estimator, achieving low $\epsilon_{PEHE}$ and high TPR faster with respect to the training set size. Moreover, all models seem to perform equivalently with 5,000 samples, including TNet, for the

three scenarios. We verify how TNet's performance is severely impaired when increasing the power and/or the number of prognostic covariates, systematically underperforming the rest of the estimators, particularly with small sample regimes. This is theoretically sound since TNet does not share information between regression tasks and, therefore, does not account for similarities between them, which is essential here as they only differ in the predictive component. We also outline the strong link between the ability of a model to give more importance to predictive covariates and the performance of the model in terms of the $\epsilon_{PEHE}$.

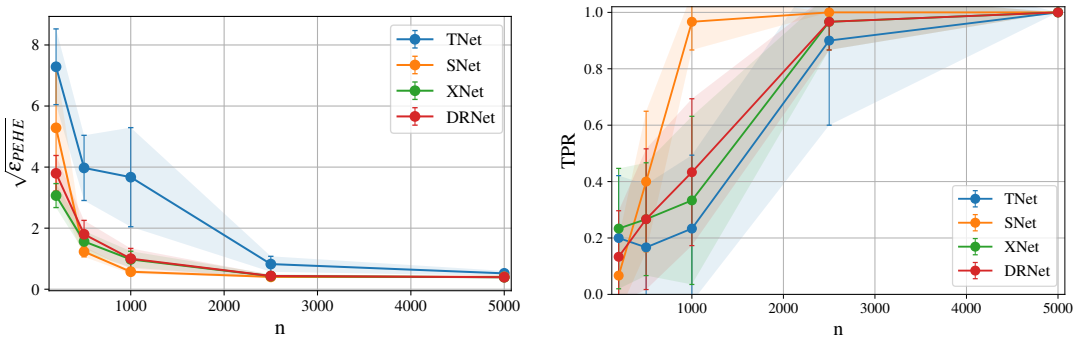



**Fig. 5.8** Performance metrics in DGP2. We train the considered CATE estimators in different sample sizes and report the mean and standard deviation of $\epsilon_{PEHE}$ and TPR for a test set with 100 samples over 10 independent simulations.

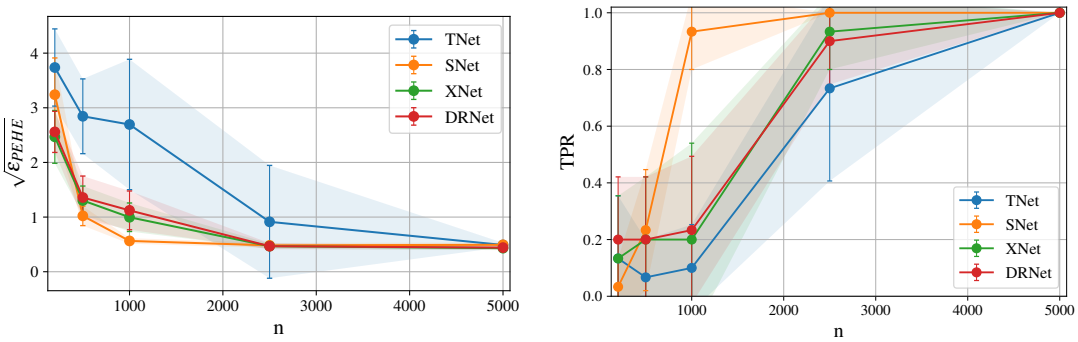
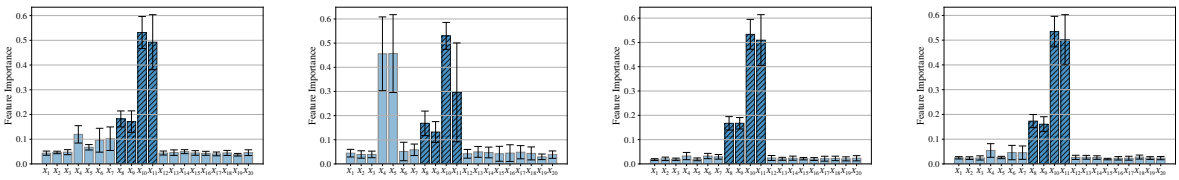
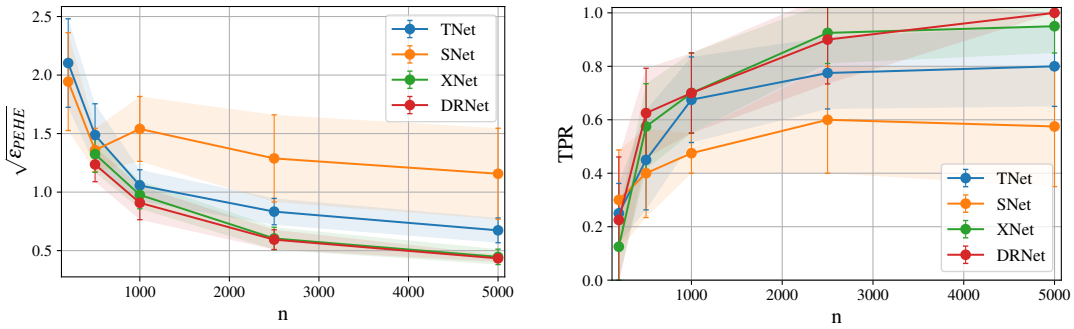



**Fig. 5.9** Performance metrics for DGP3. Same configuration as in Figure 5.8.

**More complex outcome surface** Here we include non-linear functions and interactions. Remarkably, SNet underperforms the rest of estimators for large sample regimes in DGP4 (Figure 5.10), being unable to discover the predictive information and mistakenly confounding prognostic covariates with predictive (Figure 5.10a). TNet seems to perform relatively well, and XNet and DRNet achieve a $TPR \approx 1$ for 5,000 training samples (Figure 5.10b). Figure 5.11 (DGP5) draws the same conclusions. Since it is the most complex outcome surface, XNet and DRNet need up to 10,000 samples to show a consistently good performance in $\epsilon_{PEHE}$ and TPR. SNet again shows the lowest TPR and highest $\epsilon_{PEHE}$ for 10,000 training samples.




**(a)** DGP4: $Y = 2 + 3\left(X_2 + X_3 + X_4\left(X_5 + X_6 X_7\right)\right) + W\left[\mathbb{I}\left(X_8 > 0 \cap X_9 > 0\right) + X_{10} + X_{11}^2\right] + \epsilon$




**(b)** Performance metrics $\epsilon_{PEHE}$ and TPR.

**Fig. 5.10** Results from DGP4. (a) Global feature importance given by SHAP for the CATE estimators when trained with 2,500 samples. (b) Performance metrics – $\epsilon_{PEHE}$ and TPR – for different sample sizes.

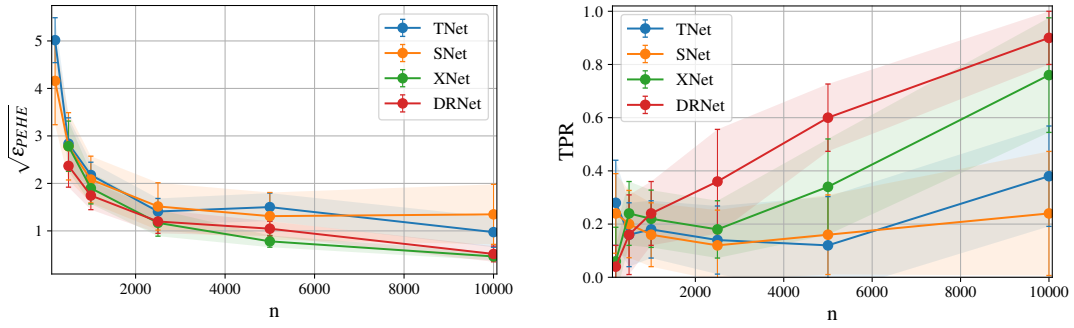



**Fig. 5.11** Performance metrics for DGP5. Same configuration as in Figure 5.8

**Increasing the treatment effect in the population** SNet is more robust than the rest of estimators against a substantial increase in the population treatment effect (Figure 5.12). All the models experiment an increase in the $\epsilon_{PEHE}$ and a decrease in the TPR.

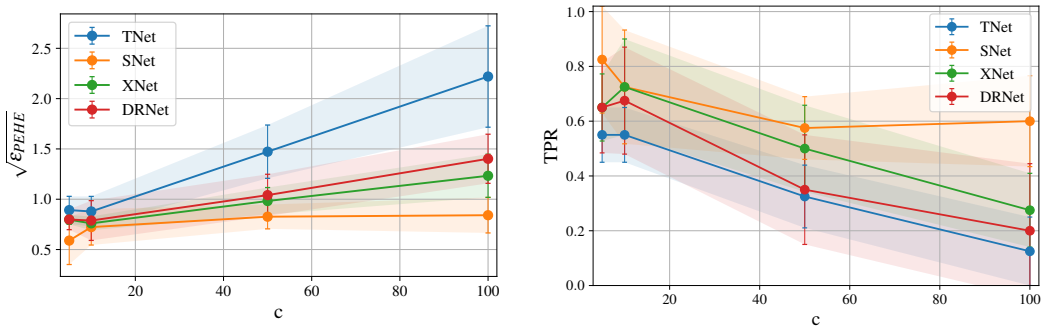
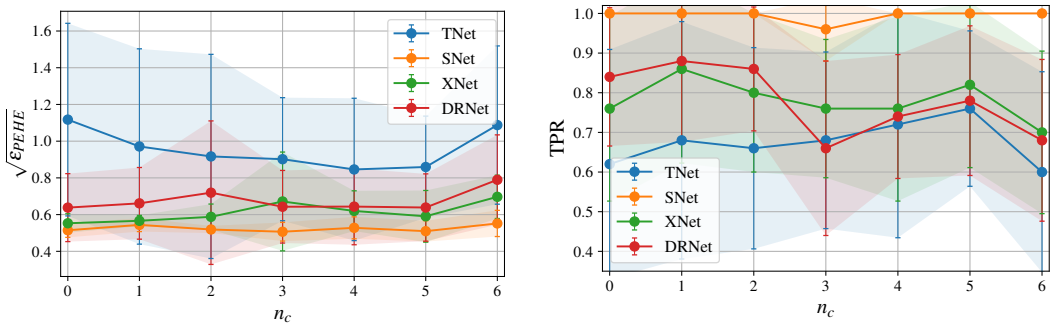
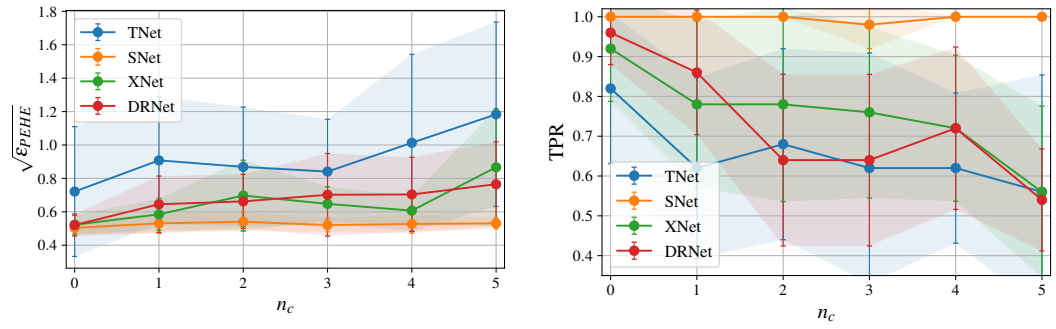



**Fig. 5.12** Performance metrics for DGP6. We train the considered CATE estimators with 2,500 samples and study the effect of increasing the overall treatment effect in the population $c$.

**Including confounders ($n_c$)** We experiment with two values for the slope of the logistic curve that determines the treatment assignment: $\xi = 2$ (Figure 5.13a) and $\xi = 3$ (Figure 5.13b). In both scenarios, SNet seems to be very robust to the covariate shift generated by confounding factors, while the rest of estimators experiment a decrease in TPR and an increase in $\epsilon_{PEHE}$. This can be justified by the enforced balanced representation in SNet, which attempts to alleviate the selection bias and, therefore, the covariate shift.



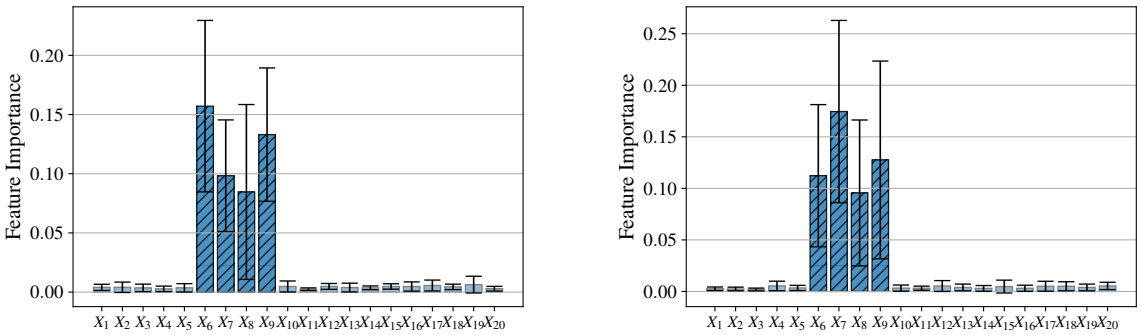

**(a)** $\xi = 2$




**(b)** $\xi = 3$

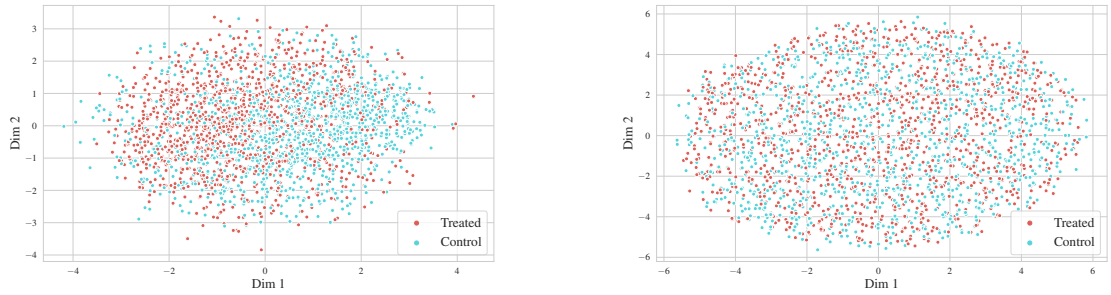
**Fig. 5.13** Performance metrics for DGP7. We train the considered CATE estimators with 2,500 samples and study the effect of increasing the number of confounders $n_c$. We consider two slopes of the logistic curve: (a) $\xi = 2$ and (b) $\xi = 3$.

It is particularly interesting to check whether the SNet features described in Section 4.2.4 are helpful in practice when selection bias exists. In Figure 5.14a, we verify that the propensity head successfully discovers the covariates that are predictive of treatment assignment (i.e., the confounders $X_6$, $X_7$, $X_8$, $X_9$) for both slopes of the logistic curve. Furthermore, Figure 5.14b illustrates how the disentanglement[2] and the imbalance loss successfully enforce a balanced representation of the factors that determine the outcome, i.e., $\Phi_O$. The t-SNE (Van der Maaten and Hinton, 2008) visualization shows the apparent imbalance between treatment groups in the original covariate space and the more balanced learned embedding $\Phi_O$. Similar results for $\Phi_{Y^0}$ and $\Phi_{Y^1}$ can be found in Appendix B.3.




$\xi = 2$ $\xi = 3$

**(a)** Saliencies obtained from the propensity head in DGP7 with $n_c = 4$. Here $X_6$, $X_7$, $X_8$, $X_9$ are confounders and thus determine the treatment assignment. SNet successfully disentangles them as relevant covariates for predicting the propensity score for $\xi = 2$ and $\xi = 3$.




Original covariate space Learned embedding $\Phi_O$

**(b)** T-SNE visualizations of treatment and control group for DGP7 with $n_c = 4$ and $\xi = 3$. The left Figure represents the units in the original covariate space, while the right Figure shows the representation learned for $\Phi_O$ by enforcing balanced representativeness of treatment groups.

**Fig. 5.14** Testing SNet features in DGP7.

[2]Figure B.4 in Appendix B.3 shows how SNet successfully disentangles the covariates following the graphical model in Figure 4.1b.

# Chapter 6

# Conclusions

## 6.1 Summary

In this thesis, we introduced a novel framework for interpreting the CATE estimation problem. The proposed methods allow discovery of the drivers of individual-level treatment effects; that is, the specific features that determine how a treatment affects a particular subject. This information is of the utmost relevance in precision medicine, where the goal is designing personalized policies to provide the clinician – and the patient – with actionable guidelines on how to benefit the most from a therapy or medication.

After describing the theoretical underpinnings in Chapter 2, Chapter 3 focused on saliency methods as the interpretability tool for explaining CATE through feature importance. We considered four well-known saliency methods – LIME, DeepLIFT, Integrated Gradients and SHAP – and we described them with a unified notation that links to the CATE setting.

Then Chapter 4 formalized what we believe makes interpretability in CATE unique and, therefore, different from the standard supervised learning paradigm. We argued that, unlike in the typical prediction setting, here we are not interested in interpreting a single (potential) outcome but the *difference* between outcomes under different treatments. We outlined that the principal aim of interpreting CATE should be disentangling the predictive covariates, which are responsible for treatment-effect heterogeneity, and we highlighted how saliency methods could be utilized to accomplish this goal. Next, we detailed the four CATE estimators that we employed in Chapter 5 to assess the framework's potential. We additionally extended a state-of-the-art architecture – SNet – by enforcing a balanced representation of the treatment groups to alleviate selection bias. Chapter 4 ended with a re-axiomatization of the elegant properties of the saliency methods by considering their utility in the CATE context.

Finally, in Chapter 5, we conducted an extensive evaluation of the framework. We started by comparing the saliency methods quantitatively and concluding that, given a sufficiently robust estimator and enough training data, the saliency methods discover the predictive covariates by identifying the most relevant features for the CATE estimation. We then illustrated their practicality in a semi-synthetic and a real-world experiments, where we investigated how their elegant properties can be used to interpret CATE. We also evidenced a lack of robustness in LIME's explanations when the CATE estimator was trained in a small-sample regime. Next, we further demonstrated the applicability of saliency methods by characterizing four CATE estimators based on their ability to disentangle predictive features in a range of synthetic experiments inspired by the drug development setting. We confirmed a consistent underperformance of TNet compared to approaches that share information between regression tasks. SNet showed systematically better results than the rest of estimators, demonstrating higher sample efficiency and robustness to selection bias, but its performance was severely impaired when dealing with highly non-linear regression surfaces. The direct estimators – XNet and DRNet – exhibited good behavior in the absence of selection bias and significant population effects, including with non-linear surfaces, but required more training samples than SNet.

Moreover, we believe that the adopted metrics and DGPs are an implicit contribution of the proposed framework. Evaluating models in causal inference problems is naturally challenging due to the absence of counterfactual information, and the proposed metrics and considered synthetic settings are sufficiently theoretically sound for a preliminary assessment of a CATE estimator, both from the predictive accuracy and the interpretability perspectives.

## 6.2 Future work

We believe that the presented framework can help advance the research on interpretability for CATE estimation; nonetheless, our developments have only scratched the surface of the problem and its potential. Several research directions could complement the present work.

First, we evaluated a relatively limited number of CATE estimators. A straightforward extension would further characterize other non-parametric approaches that include features not considered in our experiments, such as weighting schemes or more complex regularization techniques. Extending the framework for interpreting settings with multiple treatments is another exciting next step. The absence of ground truth in real data makes model selection challenging in the CATE context, and our method provides another axis of comparison between models based on their interpretability.

Furthermore, although we extensively evaluated the robustness of the framework at interpreting CATE in synthetic experiments, artificially generated data does not capture the

complexities of the real world. This was further verified in the Twins experiment, where our conclusions were less solid, mainly due to the lack of ground truth saliencies. One exciting extension is learning the underlying response surfaces from real data using AutoML-based algorithms. For example, AutoPrognosis (Alaa and van der Schaar, 2018a) can discover the risk equations from EHRs using a Bayesian Optimisation algorithm and a symbolic regression. In a recent publication (Alaa et al., 2021), the authors demonstrated the great potential of AutoPrognosis by developing and validating a treatment effect estimator – Adjutorium – based on the cancer-specific mortality risk equations obtained from the clinical profiles of nearly one million women. Moreover, integrating data from both RCTs and EHRs is also a promising avenue for future work, not only for our framework but for the CATE setting in general.

# References


Aas, K., Jullum, M., and Løland, A. (2019). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*.

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Alaa, A. and van der Schaar, M. (2018a). Autoprognosis: Automated clinical prognostic modeling via bayesian optimization with structured kernel learning. In *International conference on machine learning*, pages 139–148. PMLR.

Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J., and van der Schaar, M. (2021). Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence*, pages 1–11.

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. *arXiv preprint arXiv:1704.02801*.

Alaa, A. M. and van der Schaar, M. (2018b). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138. PMLR.

Alaa, A. M. and van der Schaar, M. (2018c). Prognostication and risk factors for cystic fibrosis via automated machine learning. *Scientific reports*, 8(1):1–19.

Alaa, A. M., Weisz, M., and Van Der Schaar, M. (2017). Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*.

Alemayehu, D., Chen, Y., and Markatou, M. (2018). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical methods in medical research*, 27(12):3658–3678.

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23(2016):139–159.

## References


Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athreya, A. P., Neavin, D., Carrillo-Roa, T., Skime, M., Biernacka, J., Frye, M. A., Rush, A. J., Wang, L., Binder, E. B., Iyer, R. K., et al. (2019). Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. *Clinical Pharmacology & Therapeutics*, 106(4):855–865.

Aumann, R. J. and Shapley, L. S. (1974). *Values of non-atomic games*. Princeton University Press.

Ballman, K. V. (2015). Biomarker: predictive or prognostic? *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 33(33):3968–3971.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bica, I., Alaa, A. M., Lambert, C., and van der Schaar, M. (2021). From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100.

Boz, O. (2002). Extracting decision trees from trained neural networks. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 456–461.

Brooks-Gunn, J., Liaw, F.-r., and Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359.

Caron, A., Manolopoulou, I., and Baio, G. (2020). Estimating individual treatment effects using non-parametric regression models: a review. *arXiv preprint arXiv:2009.06472*.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). 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*, pages 1721–1730.

Cooper, G. F., Abraham, V., Aliferis, C. F., Aronis, J. M., Buchanan, B. G., Caruana, R., Fine, M. J., Janosky, J. E., Livingston, G., Mitchell, T., et al. (2005). Predicting dire outcomes of patients with community acquired pneumonia. *Journal of biomedical informatics*, 38(5):347–366.

Crabbé, J. and van der Schaar, M. (2021). Explaining time series predictions with dynamic masks. *arXiv preprint arXiv:2106.05303*.

Curth, A. and van der Schaar, M. (2021a). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR.

Curth, A. and van der Schaar, M. (2021b). On inductive biases for heterogeneous treatment effect estimation. *arXiv preprint arXiv:2106.03765*.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, pages 598–617. Institute of Electrical and Electronics Engineers Inc.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, pages 1–12.

Foster, J., Taylor, J., and Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30 24:2867–80.

Friedman, E. J. (2004). Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4):501–518.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5.

Hahn, P. R., Carvalho, C. M., Puelz, D., He, J., et al. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

Hassanpour, N. and Greiner, R. (2019a). Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887.

Hassanpour, N. and Greiner, R. (2019b). Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Holzinger, A., Dehmer, M., and Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, 15(6):1–9.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Janzing, D., Minorics, L., and Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In *23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2907–2916. PMLR.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4).

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. (2018). Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. (2020). Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*.

Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of statistics in behavioral science*.

Julious, S. A. and Mullee, M. A. (1994). Confounding and simpson's paradox. *Bmj*, 309(6967):1480–1481.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv: Statistics Theory*.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

Kim, C. and Bastani, O. (2019). Learning interpretable models with causal guarantees. *arXiv preprint arXiv:1901.08576*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kuang, K., Cui, P., Li, B., Jiang, M., Yang, S., and Wang, F. (2017). Treatment effect estimation with data-driven variable decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Laugel, T., Renard, X., Lesot, M.-J., Marsala, C., and Detyniecki, M. (2018). Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv:1806.07498*.

Lipkovich, I., Dmitrienko, A., and B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine*, 36(1):136–196.

Loh, W.-Y., Cao, L., and Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(5):e1326.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*.

Lu, M., Sadiq, S., Feaster, D. J., and Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). Introduction to information retrieval? cambridge university press 2008. *Ch*, 20:405–416.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*.

McCauley, M. and Darbar, D. (2016). A new paradigm for predicting risk of torsades de pointes during drug development: Commentary on:"improved prediction of drug-induced torsades de pointes through simulations of dynamics and machine learning algorithms". *Clinical Pharmacology & Therapeutics*, 100(4):324–326.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582.

Moraffah, R., Karami, M., Guo, R., Raglin, A., and Liu, H. (2020). Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Nalejska, E., Mkaczyńska, E., and Lewandowska, M. A. (2014). Prognostic and predictive biomarkers: tools in personalized oncology. *Molecular diagnosis & therapy*, 18(3):273–284.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Roese, N. J. (1997). Counterfactual thinking. *Psychological bulletin*, 121(1):133.

Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Schafer, A. (1982). The ethics of the randomized clinical trial. *New England Journal of Medicine*, 307(12):719–724.

Scott, D., Bauer, C., Kraemer, H., and Tyson, J. (1989). A neonatal health index for preterm infants. In *Pediatric Research*, volume 25, pages A263–A263. WILLIAMS & WILKINS 351 WEST CAMDEN ST, BALTIMORE, MD 21201-2436.

Sechidis, K., Papangelou, K., Metcalfe, P. D., Svensson, D., Weatherall, J., and Brown, G. (2018). Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19):3365–3376.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Shapley, L. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Shapley, L. S. and Shubik, M. (1971). The assignment game i: The core. *International Journal of game theory*, 1(1):111–130.

Shi, C., Blei, D. M., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Sotoudeh, M. and Thakur, A. V. (2019). Computing linear restrictions of neural networks. *arXiv preprint arXiv:1908.06214*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599.

Sturmfels, P., Lundberg, S., and Lee, S.-I. (2020). Visualizing the impact of feature attribution baselines. *Distill*. https://distill.pub/2020/attribution-baselines.

Sundararajan, M. and Najmi, A. (2019). The many Shapley values for model explanation. *arXiv*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Svensson, D. and Hermansson, E. (2021). On discovering treatment-effect modifiers using virtual twins and causal forest ml in the presence of prognostic biomarkers.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wu, A., Kuang, K., Yuan, J., Li, B., Zhou, P., Tao, J., Zhu, Q., Zhuang, Y., and Wu, F. (2020). Learning decomposed representation for counterfactual inference. *arXiv preprint arXiv:2006.07040*.

Xu, G., Duong, T. D., Li, Q., Liu, S., and Wang, X. (2020). Causality learning: A new perspective for interpretable machine learning. *arXiv preprint arXiv:2006.16789*.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2020). A survey on causal inference. *arXiv preprint arXiv:2002.02770*.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2019). Ace: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1432–1437. IEEE.

Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Zhang, Z., Seibold, H., Vettore, M. V., Song, W.-J., and François, V. (2018). Subgroup identification in clinical trials: an overview of available methods and their implementations with r. *Annals of translational medicine*, 6(7).

# Appendix A

# Mathematical Derivations

## A.1 Maximum Mean Discrepancy empirical estimate

Let $p$ and $q$ be two probability density functions, with i.i.d observations $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_n\}$ on a topological space $\mathcal{R}$, so that $x \sim p$ and $y \sim q$. From the IPM expression in Equation (4.11), given a class of functions $\mathscr{F}$ with $f : \mathcal{R} \to \mathbb{R}$, we define

$$\mathrm{IPM}_{\mathscr{F}}(p, q) := \sup_{f \in \mathscr{F}} \left( \mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)] \right). \tag{A.1}$$

In the MMD, the function class $\mathscr{F}$ is the unit ball in a reproducing kernel Hilbert space $\mathcal{H}$. Therefore:

$$\mathrm{MMD}^2(p, q) = \left[ \sup_{\|f\|_{\mathcal{H}} \leq 1} \left( \mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)] \right) \right]^2 \tag{A.2}$$

$$= \left[ \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} \right]^2 \tag{A.3}$$

$$= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 . \tag{A.4}$$

Making the expansion:

$$\begin{aligned} \mathrm{MMD}^2(p, q) &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \\ &= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2 \langle \mu_p, \mu_q \rangle_{\mathcal{H}} \qquad \text{(A.5)} \\ &= \mathbb{E}_{x, x' \sim p} \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} + \mathbb{E}_{y, y' \sim q} \langle \phi(y), \phi(y') \rangle_{\mathcal{H}} - 2 \mathbb{E}_{x \sim p, y \sim q} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}. \qquad \text{(A.6)} \end{aligned}$$

Finally, by applying the kernel trick and estimating the population expectations with their corresponding U-statistics and sample averages (Gretton et al., 2012):

$$\mathrm{MMD}^2(p,q) = \mathbb{E}_{x,x'\sim p}\left[k\left(x,x'\right)\right] + \mathbb{E}_{y,y'\sim q}\left[k\left(y,y'\right)\right] - 2\mathbb{E}_{x\sim p,y\sim q}[k(x,y)] \tag{A.7}$$

$$\approx \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k\left(x_i,x_j\right) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k\left(y_i,y_j\right)$$

$$- \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k\left(x_i,y_j\right). \tag{A.8}$$

This statistic is unbiased following (Serfling, 2009).

# Appendix B

# Additional Experiments

## B.1 Tuning SNet: imbalance loss term

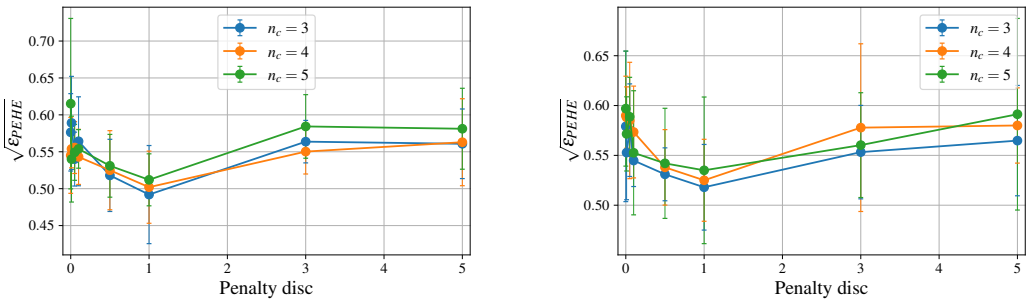



**Fig. B.1** Tuning penalty $\beta$ in SNet. Note that with $\beta = 1$ the $\epsilon_{PEHE}$ improves (decreases) on average compared to not including an imbalance loss (i.e., $\beta = 0$) for our experiments. These simulations correspond to DGP7 in Section 5.3.2 with (a) $\xi = 2$ and (b) $\xi = 3$.

## B.2 Consistency checks for saliency methods

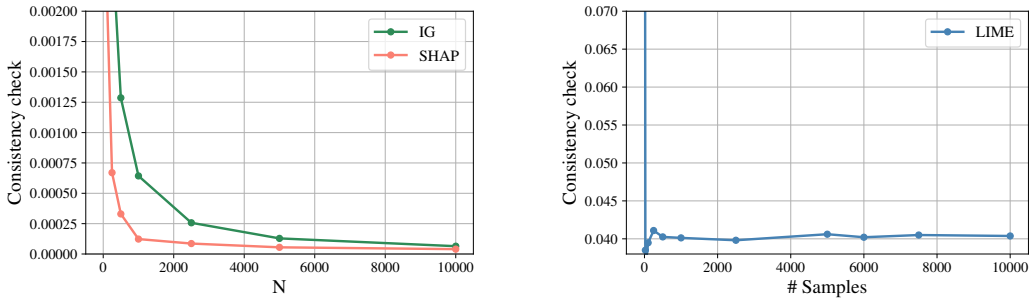
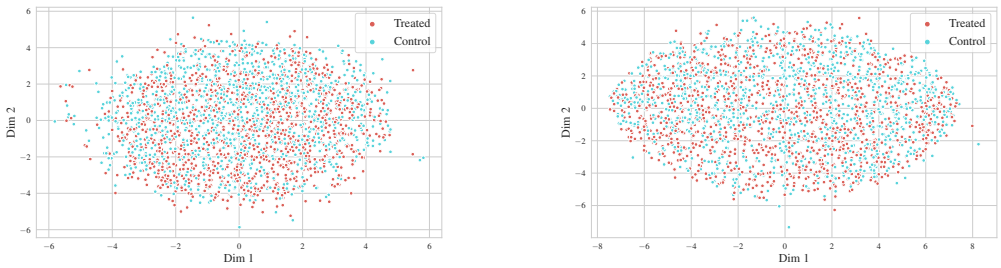



**Fig. B.2** Consistency checks for selecting the number of step in approximating the IG Riemann sum, the number of samples drawn for KernelSHAP, and the number of samples to train the interpretable model in LIME. In IG and SHAP we use the completeness property $|\sum_{c=1}^{d} a_c(\hat{\tau}, \boldsymbol{x}) - \hat{\tau}(\boldsymbol{x}) + b|$, and for LIME the local accuracy $|\sum_{c=1}^{d} a_c(\hat{\tau}, \boldsymbol{x}) - \hat{\tau}(\boldsymbol{x})|$. This simulation was ran in the IHDP experiment for illustration, but similar checks were conducted for all the experiments.

## B.3 Other results




(a) (b)

**Fig. B.3** T-SNE visualizations of the learned embeddings (a) $\Phi_{Y^0}$ and (b) $\Phi_{Y^1}$ in DGP7.

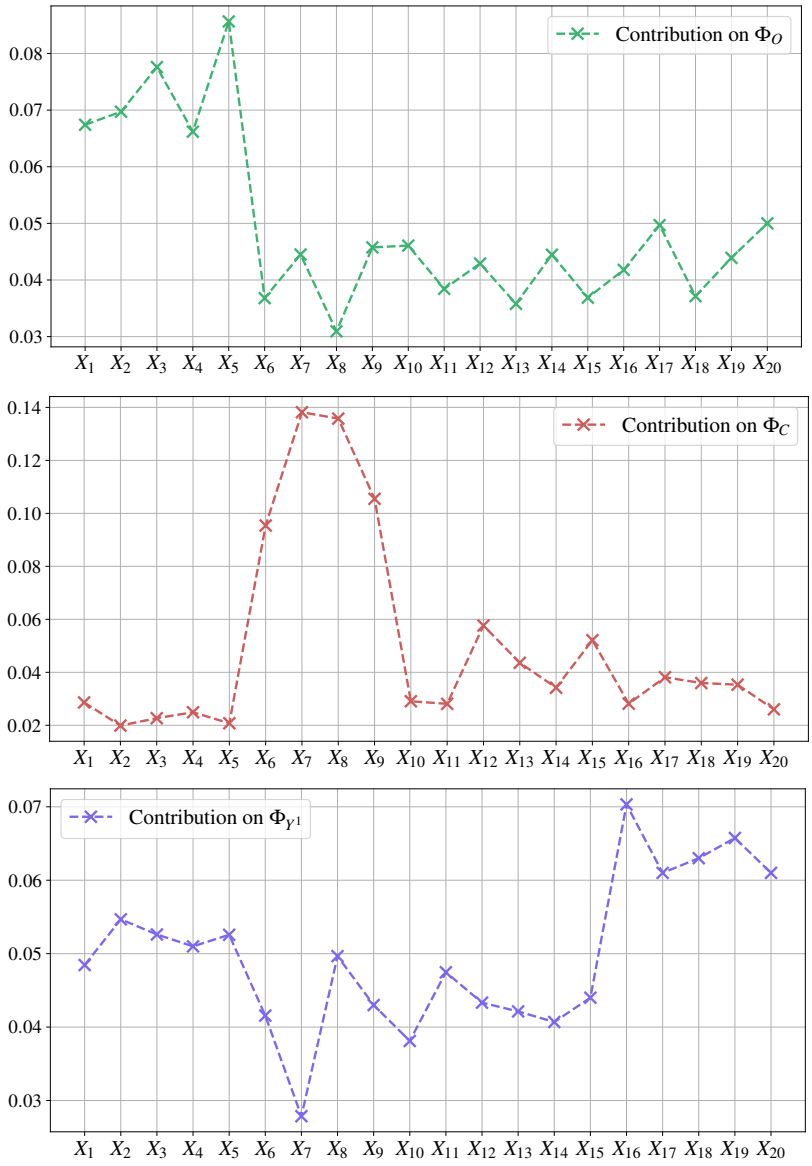



**Fig. B.4** Contribution of each covariate in the decomposed representations $\Phi_O$, $\Phi_C$ and $\Phi_{Y^1}$ for DGP7 with $n_c = 4$ and $\xi = 3$. Recall that the contribution of a covariate $j$ in representation $\Phi_k$ is approximated by $\bar{W}_{\Phi_k,j} = \sum_u \left|\boldsymbol{W}_{j,u}^{1,\Phi_k}\right|$, normalized here so that the contribution of all covariates sums to one for each representation. Note that the disentanglement is successful, having the prognostic covariates $\boldsymbol{X}_Y = \{X_1, X_2, X_3, X_4, X_5\}$ the largest contribution in representation $\Phi_O$, the confounders $\boldsymbol{X}_C = \{X_6, X_7, X_8, X_9\}$ in $\Phi_C$, and the predictive covariates $\boldsymbol{X}_{Y^1} = \{X_{16}, X_{17}, X_{18}, X_{19}, X_{20}\}$ in $\Phi_{Y^1}$. Each contribution is averaged across 10 simulations, but only the mean is shown for clarity.

# Appendix C

# Datasets Details

## C.1 IHDP dataset

| Number | Name | Description | Type |
|---|---|---|---|
| $X_1$ | bw | birth weight | continuous |
| $X_2$ | b.head | head circumference | continuous |
| $X_3$ | preterm | weeks born preterm | continuous |
| $X_4$ | birth.o | birth order | continuous |
| $X_5$ | nnhealth | neonatal health index (Scott et al., 1989) | continuous |
| $X_6$ | momage | mother's age | continuous |
| $X_7$ | sex | sex of child | binary |
| $X_8$ | twin | twin or not twin | binary |
| $X_9$ | b.marr | mother's marital status | binary |
| $X_{10}$ | mom.lths | mother did not go to high school | binary |
| $X_{11}$ | mom.hs | mother graduated from high school | binary |
| $X_{12}$ | mom.scoll | mother graduated from college | binary |
| $X_{13}$ | cig | mother smoked during pregnancy | binary |
| $X_{14}$ | first | first born | binary |
| $X_{15}$ | booze | mother drank | binary |
| $X_{16}$ | drugs | mother took drugs | binary |
| $X_{17}$ | work.dur | mother worked | binary |
| $X_{18}$ | prenatl | received prenatal care | binary |
| $X_{19}$ | ark | data collected from University of Arkansas for Medical Sciences | binary |
| $X_{20}$ | ein | Albert Einstein College of Medicine | binary |
| $X_{21}$ | har | Harvard Medical School | binary |
| $X_{22}$ | mia | University of Miami School of Medicine | binary |
| $X_{23}$ | pen | University of Pennsylvania School of Medicine | binary |
| $X_{24}$ | tex | University of Texas Health Science Center at Dallas | binary |
| $X_{25}$ | was | University of Washington School of Medicine | binary |

**Table C.1** Details on the original IHDP dataset. Non-binary covariates are previously min-max scaled, resulting in continuous values. More information in the original publication (Brooks-Gunn et al., 1992).

## C.2 Twins dataset

| Number | Name | Description | Type |
|---|---|---|---|
| $X_1$ | dmage | age mother | continuous |
| $X_2$ | mpcb | month prenatal care began | continuous |
| $X_3$ | cigar | average number of cigarettes per day | continuous |
| $X_4$ | drink | average number of drinks per week | continuous |
| $X_5$ | wtgain | weight gain | continuous |
| $X_6$ | gestat | weeks of gestation | continuous |
| $X_7$ | dmeduc | education mother | continuous |
| $X_8$ | nprevist | number of prenatal visits | continuous |
| $X_9$ | dmar | married | binary |
| $X_{10}$ | anemia | risk factor, Anemia | binary |
| $X_{11}$ | cardiac | risk factor, Cardiac | binary |
| $X_{12}$ | lung | risk factor, Lung | binary |
| $X_{13}$ | diabetes | risk factor, Diabetes | binary |
| $X_{14}$ | herpes | risk factor, Herpes | binary |
| $X_{15}$ | hydra | risk factor Hydramnios/Oligohidramnios | binary |
| $X_{16}$ | hemo | risk factor, Hemoglobinopathy | binary |
| $X_{17}$ | chyper | risk factor, Hypertension, chronic | binary |
| $X_{18}$ | phyper | risk factor, Hypertension, pregnancy-associated | binary |
| $X_{19}$ | eclamp | risk factor, Eclampsia | binary |
| $X_{20}$ | incervix | risk factor, Incompetent cervix | binary |
| $X_{21}$ | pre4000 | risk factor, Previous infant 4000+ grams | binary |
| $X_{22}$ | dtotord | total number of births before twins | binary |
| $X_{23}$ | preterm | risk factor, Previous infant pre-term or small | binary |
| $X_{24}$ | renal | risk factor, Renal disease | binary |
| $X_{25}$ | rh | risk factor, RH sensitization | binary |
| $X_{26}$ | uterine | risk factor, Uterine bleeding | binary |
| $X_{27}$ | othermr | risk factor, Other Medical Risk Factors | binary |
| $X_{28}$ | adequacy_1 | adequate adequacy | binary |
| $X_{29}$ | adequacy_2 | intermediate adequacy | binary |
| $X_{30}$ | adequacy_3 | inadequate adequacy | binary |
| $X_{31}$ | pldel_1 | born in hospital | binary |
| $X_{32}$ | pldel_2 | born in birthing center | binary |
| $X_{33}$ | pldel_3 | born in clinic or doctor's office | binary |
| $X_{34}$ | pldel_4 | born in a residence | binary |
| $X_{35}$ | pldel_5 | born in other place | binary |
| $X_{36}$ | resstatb_1 | resident in US | binary |
| $X_{37}$ | resstatb_2 | intrastate nonresident | binary |
| $X_{38}$ | resstatb_3 | interstate nonresident | binary |
| $X_{39}$ | resstatb_4 | foreign resident | binary |

**Table C.2** Details on the Twins dataset. Non-binary covariates are previously min-max scaled, resulting in continuous values. We obtain the dataset from (Yoon et al., 2018). More information in the original publication (Almond et al., 2005)

## C.3 Data Generation Processes

All covariates are sampled from a multivariate Gaussian with zero mean and arbitrary covariance matrix, ensuring $\Sigma_{ij} < 0.3$ between them. Selection bias is induced in experiments with confounders $\boldsymbol{X_C}$ and/or with covariates predictive of treatment assignment $\boldsymbol{X_W}$. Here treatments are assigned as $W \mid \boldsymbol{X} \sim \text{Bern}\left(\sigma\left(\xi \cdot [\boldsymbol{X_C}, \boldsymbol{X_W}]\ \mathbf{w} + n\right)\right)$, where $\mathbf{w} \sim \mathcal{N}\left(0, \mathbf{I}^{(n_c+n_w)\times 1}\right)$ and $n \sim \mathcal{N}(0, 0.1)$. $\xi$ is the slope of the logistic curve. In the RCT scenario there is not selection bias so treatments are assigned at random: $W \sim \text{Bern}(0.5)$.