IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 January 2024, accepted 7 March 2024, date of publication 11 March 2024, date of current version 18 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376423

RESEARCH ARTICLE

Undersmoothing Causal Estimators With Generative Trees

DAMIAN MACHLANSKI[1], **SPYRIDON SAMOTHRAKIS**[2], **AND PAUL CLARKE**[3]

[1]Department of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K.
[2]Institute for Analytics and Data Science, University of Essex, CO4 3SQ Colchester, U.K.
[3]Institute for Social and Economic Research, University of Essex, CO4 3SQ Colchester, U.K.

Corresponding author: Damian Machlanski (d.machlanski@essex.ac.uk)

This work was supported in part by the Economic and Social Research Council (ESRC) Research Centre on Micro-Social Change (MiSoC) under Grant ES/S012486/1, and in part by the computational resources provided by the Business and Local Government Data Research Centre (BLG DRC) funded by ESRC under Grant ES/S007156/1.

ABSTRACT Average causal effects are averages of (heterogeneous) individual treatment effects (ITEs) taken over the entire target population. The estimation of average causal effects has been studied in depth, but averages are insufficient for more individualised decision-making where ITEs are more appropriate. However, estimating ITEs for every population member is challenging, particularly when estimation must be based on observational data rather than data from randomised experiments. One potential problem with observational data arises when there are large differences between the sample distributions of the input features of the treated and control units. This problem is known as covariate shift. It can lead to model misspecification the harmful effects of which can be severe for ITE estimation because point estimation is highly sensitive to regions of the common support of the input space in which the number of treated or control units is very small. Moreover, common solutions are often based on reweighing schemes involving propensity scores which were originally designed for average effects and not ITEs. In this paper, we propose Debiasing Generative Trees, a novel data augmentation method based on generative trees that debiases and undersmooths causal estimators trained on augmented data. It encourages higher modelling complexity that reduces misspecification and improves estimation of ITEs. We show empirically that our proposed approach yields models of higher complexity and more accurate predictions of ITEs, and is competitive with traditional methods for estimating average treatment effects. Our results confirm that reweighing methods can struggle with ITE estimation and that the choice of model class can significantly impact prediction performance.

INDEX TERMS Conditional average treatment effect, observational data, covariate shift, model misspecification, data augmentation, generative trees.

I. INTRODUCTION

In the absence of data from randomised experiments, analysts must use observational data to make inferences about the causal effects of interventions or treatments, that is, what would happen if they intervened to change the treatment status of individual units in a population. The estimation of average causal effects — the average effect of the treatment aggregated across every unit in a population — has been studied in considerable depth. However, there is now growing interest in estimating heterogeneous treatment effects for individuals characterized by a possibly large number of input variables or covariates. If there is substantial heterogeneity across units, such systems can unlock the analysis of targeted interventions, for instance, in the form of personalised healthcare based on patients' symptoms and health histories.

The use of observational data creates challenges for the estimation of heterogeneous causal effects. First, the analyst must make assumptions, for example, that treatment selection is strongly ignorable given the available covariates. We take ignorability to hold throughout, and focus on the second problem, namely, that nonrandom treatment selection can lead to observed data in which the distributions of covariates among the treated and untreated units are very different. In practice, this can make it difficult for conventional learners to learn the true relationship between the treatment effect and

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

covariates across the entire support of the covariates, and so result in poor performance when tested on other datasets.

More generally, this issue is known as ‘covariate shift’, which in this setting means the learning target $P(Y|X)$ remains unchanged, while the marginal distributions of the covariate inputs $P(X)$ for treated and untreated can be very different. Most existing methods attempt to transform the observational distribution by sample reweighing schemes usually based on propensity scores [1], [2], [3], [4], [5] (but not exclusively, see e.g. domain adaptation methods). However, reweighing seeks to standardise the observed support of X for the treated and untreated groups, and so generally performs well for estimating treatment effects averaged across the common support of X, but less so for estimating conditional average treatment effects at points outside the observed support; in other words, as pointed out by [6], reweighing does not address the problem of model misspecification which can be detrimental when it comes to estimating individualised treatment effects [7].

A promising alternative to these classical approaches is undersmoothing, where the model is allowed to fit the data very closely to capture $P(X)$ in the two treatment groups, and in doing so potentially produce more accurate individualised predictions. Encouraged by suggestions elsewhere - [8, footnote 3] and [9], [10] - in this paper, we develop a novel approach to causal effect estimation that improves accuracy by undersmoothing the observed data.

Specifically, we propose to undersmooth using fast and straightforward generative trees [11] to augment the existing data, and in doing so facilitate more robust learning of downstream estimators of key causal parameters. The trees are used to ‘discretise’ the input space into subpopulations of similar units (subclassification); the distributions of these groups are then modelled separately via mixtures of Gaussians, from which we sample equally to reduce data biases.

The concept of model misspecification comes from the world of finite-dimensional parametric models [7] when the analyst uses a parametric model family for prediction that does not include the true prediction rule implied by the Data Generating Process (DGP). In our context, where no such family is specified, the data augmentation algorithm leads to individualised predictions which can be viewed as coming from an infinite-dimensional model family for statistical functionals that, while not nonparametric, is richer than those induced by existing alternative algorithms. We argue that the implied model family is more likely to include the DGP because data augmentation oversampling the underrepresented data regions is effectively performing targeted undersmoothing and so reduces bias [9]. The practical upshot of this should be that, even with covariate shift, learners trained on data augmented by our method offer more accurate predictions.

Data augmentation is a widely recognised data preprocessing method of improving overall data quality through synthetic sample generation for better prediction performance [12]. It has proven very effective in computer vision [13], [14] which greatly influenced wider popularisation of data augmentation techniques. Data augmentation constitutes an important part of dealing with imbalanced tabular data [15], specifically by oversampling minority classes in imbalanced classification problems [16], [17], [18], [19]. Interestingly, causal notions and ability to simulate interventions has been attributed to successes of data augmentation [20]. This links to a specific type of augmentation that focuses on generating counterfactuals (unobserved outcomes), called counterfactual data augmentation, which found its use in classification problems [21], [22] and reinforcement learning [23]. Other methods focus on text classification [24] or mitigating the effects of confounding [25]. In our case, the method we propose could be seen as oversampling underrepresented data regions instead of just classes or specific outcomes like counterfactuals, making our approach much more general.

Generative models have also been investigated in causal inference literature, mostly in two major strands of work. In one, generative models are used for benchmarking purposes to create new synthetic data sets that closely resemble real data but with access to true, though synthetically generated, effects [26], [27]. The other branch of research is concerned with generating causal effects [28], with more recent works applied to bounding confounded average effects [29], continuous treatments under confounding [30], and longitudinal data [31]. In this work, unlike the two strands above, we use generative modelling for targeted data augmentation.

Arguably the closest work to ours that combines data augmentation and generative models within the causal inference setting is [32]. Despite a similar approach on a high-level, that is, train downstream causal estimators on augmented data, we believe our frameworks differ substantially upon further examination. More precisely, [32] incorporates neural network based generative models to specifically generate counterfactuals and focuses on conditions where the treatment is continuous. In this work, our proposed method: a) is based on simple and widely-used decision trees, b) does not specifically generate counterfactuals, but oversamples heterogeneous data regions (more general), and c) works with classic discrete treatments.

In terms of this paper’s contributions, we show empirically that the choice of model class can have a substantial effect on estimator’s final performance, and that standard reweighing methods can struggle with individual treatment effect estimation. Given our experiments, we also provide an evidence that our proposed method increases data complexity, reduces bias by training on augmented data (targeted undersmoothing), and leads to statistically significant improvements in individual treatment effect estimation, while keeping the average effect predictions competitive. Our experimental setup incorporates a wide breadth of non-neural standard causal inference methods and data sets. We specifically focus on non-neural solutions as they are more commonly used by practitioners.

The rest of the document is structured as follows. In Section II, we revisit fundamental concepts that should aid understanding of the technical part of the paper. Next, we formally discuss the problem of model misspecification (Section III), followed by a thorough description of our proposed method in Section IV. We then present our experimental setup and results (Section V). Next section provides further discussion on the results (Section VI), and considered limitations of the method (Section VII). Section VIII concludes the paper.

II. PRELIMINARIES

This section gives a brief overview of the essential background deemed relevant to this work. For a more extensive review, we refer the reader to classic positions on causal analysis [33], [34], and recent surveys on causal inference [35], [36].

A. TREATMENT EFFECT ESTIMATION

Given two random variables T and Y, investigating effects of interventions can be described as measuring how the outcome Y differs across different inputs T. Real-world systems usually contain other background covariates, denoted as X, which have to be accounted for in the analysis as well. To formally approach the task, we take Rubin's Potential Outcomes [37] perspective, which is particularly convenient in outcome estimation without knowing the full causal graph.

We start by defining the potential outcomes $\mathcal{Y}_t^{(i)}$, that is, the observed outcome when individual i receives treatment $t = 0, 1$. Given this, the Individual Treatment Effect (ITE) is formulated as the difference between the outcomes under treatment ($\mathcal{Y}_1^{(i)}$) and no treatment ($\mathcal{Y}_0^{(i)}$), as in (1).

$$\text{ITE}_i = \mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}. \tag{1}$$

Thus, to compute such a value for individual i, we need access to both potential outcomes, $\mathcal{Y}_1^{(i)}$ and $\mathcal{Y}_0^{(i)}$, but only one, called the *factual*, is observed: the other potential outcome, called the *counterfactual*, cannot be observed. The fact that we only observe factuals but also need the counterfactuals to properly compute causal effects is known as the fundamental problem of causal inference: ITEs are not identified by the observed data.

However, parameters such as the Average Treatment Effect (ATE) and Conditional Average Treatment Effect (CATE) are identified, defined in (2) and (3) respectively.

$$\text{ATE} = \mathbb{E}\left[\mathcal{Y}_1 - \mathcal{Y}_0\right] \tag{2}$$
$$\text{CATE} = \mathbb{E}\left[\mathcal{Y}_1 | X = x\right] - \mathbb{E}\left[\mathcal{Y}_0 | X = x\right], \tag{3}$$

where $\mathbb{E}[.]$ denotes (statistical) expectation over the target population. The ATE in (2) is essentially the average ITE for the entire population (hence the expectation operator and no index i); the CATE in (3) is the average ITE for everyone in the subpopulation characterised by $X = x$. The ATE is not meaningful if there is substantial heterogeneity of the ITEs between subpopulations. In such circumstances, CATE is more informative about ITEs as it allows the effect to be conditioned on the subpopulation of interest. The ITE can be thought of as a special case of CATE where individual i is the only member of the subpopulation. While ITE_i cannot be identified, CATE for the subpopulation $X = x$ which includes individual i will be better estimate of it than ATE (under the reasonable assumption that between-subpopulation variation in ITEs is greater than that within subpopulations).

B. COMMON ASSUMPTIONS

When it comes to causal effect estimation, there are three common assumptions about the data generating process that many estimators build upon. These are: SUTVA, ignorability and positivity. As we assume the three throughout the paper, we include their brief description for completeness.

Assumption 1 (Stable Unit Treatment Value Assumption (SUTVA)): The potential outcomes of any unit (individual) do not affect the treatment assignment of any other unit. Furthermore, there are no different levels or forms of the same treatment. In short: a) no inter-unit interaction, and b) no hidden treatment variations.

Assumption 2(Ignorability): Given background covariates X, potential outcomes $\mathcal{Y}$ are independent from observed treatment T. That is, $\mathcal{Y}_1, \mathcal{Y}_0 \perp\!\!\!\perp T | X$. In practice, it means that for individuals with the same X their treatment assignment T can be perceived as random because there are no unmeasured hidden variables (confounders), which is why this assumption is often referred to as *unconfoundedness*.

Assumption 3 (Positivity): Treatment assignment T is not deterministic for all individuals X. That is, $P(T = t|X = x) > 0$ for all t and x. Informally, it requires the existence of potential outcomes for all treatments and individuals (and their combinations). Ignorability and positivity together form *strong ignorability*.

C. COVARIATE SHIFT

The covariate-shift problem generally occurs when there are distributional discrepancies in input variables X among certain groups of data samples. These differences lead to 'gaps', that is, regions of the common support of X where the observed data are non- or weakly informative about the target parameter. As a result, point estimates of the target (e.g. ITEs or CATEs) in these gaps are inaccurate. Note this issue is not as severe for population-level estimates (e.g. ATEs) because these are averages across the entire support of X and so more robust to gaps. The problem varies depending on the characterisation of the groups of data among which the covariate shift occurs. In Machine Learning (ML), this problem is often realised when there are differences between training and test (target) distributions. These can happen due to, for instance, different circumstances between data collection and model deployment. In causal inference, on the other hand, said distributional discrepancies exist between treated and control units. For example, in a study of smoking effects on health, the observational data at hand may include very little to no information about young smokers. Methods of dealing with covariate shift comprise data adaptation [38], importance sample reweighing (see Section II-D below), or causal effect estimation in general (see Section II-E below).

Covariate shift falls under a broader category of distribution shifts, in which a second major shift problem occurs due to changes in the conditional distribution $Y|X$ between the target and input covariates. Specific reasons behind may involve confounding but also a change in subject behaviour over time [39]. Regardless of the nature of data shifts, they have been shown to have a clear impact on prediction performance [39], [40], and the fact that real-world data sets often suffer from not one but a mixture of types of shifts only adds importance to this issue. Robustness to data shifts is also discussed from the perspective of out-of-distribution generalisation [41], a topic researched with renewed interest in ML recently partly due to still unresolved issues with model inaccuracies after deployment.

D. CLASSIC REWEIGHING

Reweighing methods seek to transform the observed support of input covariates for the treated and control groups via *propensity scores*. These are simply defined as unit's probability of receiving the treatment, that is, $e(x_i) = P(t_i|x_i)$. One straightforward way of using such scores to aid the aforementioned covariate shifts between treated and untreated units is called Inverse Propensity Weighting (IPW) [3], also known as Inverse Probability of Treatment Weighting (IPTW), which defines weight w_i for sample i that depends on treatment status t_i and propensity $e(x_i)$, as in (4).

$$w_i = \frac{t_i}{e(x_i)} + \frac{1-t_i}{1-e(x_i)}. \tag{4}$$

The weight in (4) can be also perceived as sample importance. The higher the weight, the more impact that sample should have on the estimator during training. Its inverse nature refers to the weight and probability of treatment being inversely proportional. Thus, assuming treated units are less numerous as compared to the untreated ones, this approach assigns higher weights to treated samples hence providing a balancing effect.

The main drawback of this approach is its heavy reliance on the accuracy of the propensity scores. To counteract this, Doubly Robust [2] proposes to mix IPW and outcome regression, resulting in a method robust to misspecifications of either (but not both). An alternative improvement is to use propensity scores to balance not only samples but covariates as well [42]. IPW can also result in extremely small scores, making the entire estimation very unstable. One possible solution is trimming, that is, to eliminate samples with propensity scores lower than a specified threshold [43]. Another way of improving and stabilising propensity score models is post-calibration [44].

E. MODERN ESTIMATORS

Many modern approaches to causal effect estimation continue to incorporate propensity scores in one form or another but they are not necessarily their main feature or contribution.

Perhaps the simplest and most naive approach that does not involve reweighing is regression adjustment, where a single regressor $\mu(x, t)$ is used to estimate potential outcomes from which causal effects $\tau(x)$ can be calculated, such that $\tau(x) = \mu(x, 1) - \mu(x, 0)$. Due to distributional differences between treated and untreated, using separate regressors $\mu_t(x)$ per treatment arm $t = 0, 1$ might be preferable, resulting in $\tau(x) = \mu_1(x) - \mu_0(x)$. The two approaches have been formalised as S-Learner and T-Learner respectively [4], with the authors also proposing their own X-Learner that combines the ideas of T-Learning with propensity scores $e(x)$ that control the degree of contribution of each arm's model in the final estimation step.

Further, a general parametric CATE framework was introduced as part of Double Machine Learning [1] which combines Neyman-orthogonal equations and cross-fitting to reduce estimation bias of the nuisance parameters $\mu(x)$ and $e(x)$. The R-Learner further generalises this approach to nonparametric CATEs [45], of which the X-Learner has been shown to be a special case as well. All these procedures can be categorised more generally as Orthogonal Learners [46].

Ensemble methods have also been explored in causal effect estimation. Targeted Maximum Likelihood Estimation (TMLE) realises this approach via its "super-learning" and use of influence functions [47], [48]. Causal Forests [5] build an ensemble of "honest" decision trees that are based on causal effect heterogeneity as a splitting strategy. Ensembles of small Neural Networks (NNs) also showed great promise [49].

Recent successes of neural networks also resulted in NN-specific solutions to treatment effect estimation. Many build on the basic principles of feed-forward NNs, but modify their loss functions to encourage representations of treated and control groups that are balanced [50], [51]. More advanced architectures often employ a so-called "two-head" approach wherein each NN output is dedicated to a separate treatment arm [52]. A solution with three heads was also proposed, in which the third output optimises for propensities $e(x)$ that are used for weighting purposes [53]. Interestingly though, sample importances have been shown to have negligible effect on deep networks [54]. In terms of generative NNs, both Variational Autoencoders [28] and Adversarial Networks [55] have been successfully explored as well.

The latest developments show even greater diversity in proposed methodologies and are a testament of continued research interest in the causal estimation problem. Neural architectures are still being explored in this line of work, most recently in the form of *normalising flows* that target CATE distribution modelling instead of expected values, providing uncertainty quantification as a consequence [56]. Another research front continues the developments concerned with metalearners. One important example includes the addition of the conformal prediction framework on top of metalearning that enables predictive intervals for ITEs as opposed to CATE point estimates [57]. Conformal learning indeed has been gathering increased interest in causal estimation as it found its way into offline off-policy prediction problems as well [58]. B-Learner constitutes another proposed metalearner that tackles the hidden confounding problem and provides bounds on predicted CATEs [59]. Some other notable work involves forecasting treatment outcomes over time [60] and a comparative study on performance differences between parametric and nonparametric causal effect modelling [61].

III. MODEL MISSPECIFICATION

The choice of model class occurs at some point in any learning task. Such a decision is made based on available data, usually the training part of it, while the environment of the actual application can be different, a scenario often mimicked via a separate test set. The occurring discrepancies between those two data sets are known as covariate shift problem. Within causal inference, this manifests as differences between observational and interventional distributions, ultimately making effect estimation extremely difficult. More formally, given input covariates x, treatment t, and outcome y, the conditional distribution $P(y|x, t)$ remains unchanged across the entire data set, whereas marginal distributions $P(x, t)$ differ between observational and interventional data. This is where model misspecification occurs as the model class is selected based on available observations only, which does not generalise well to later predicted interventions.

Let us consider a simple example as presented in Fig. 1. It consists of a single input feature x, output variable y (both continuous), and binary treatment t. For convenience, let us denote this data set as D. Note the effect is clearly heterogeneous as it differs in $D(x < 0.5)$ and $D(x > 0.5)$. Furthermore, the two data regions closer to the top of the figure, that is, $D(x < 0.5, t = 1)$ and $D(x > 0.5, t = 0)$, are in minority with respect to the rest of the data. By many learners these scarce data points will likely be treated as outliers, resulting in lower variance than needed to provide accurate estimates. Thus, naively fitting the data will lead to biased estimates, an example of which is depicted on the figure as *Biased T* and *Biased C*. However, what we aim for is an unbiased estimator that captures the data closely while still generalising well, a scenario showcased by *Unbiased T* and *Unbiased C* on the figure.

For ITE estimation, fitting the data closely is especially important. Although in case of average effect estimation the difference between biased and unbiased estimators can be negligible, the individualised case usually exacerbates the issue. For instance, in the presented example, the difference in ATE error is 0.44, but it grows to 0.77 in ITE error.

In this work, instead of altering the sample importance, as many existing methods do, we aim to augment provided data in a way that underrepresented data regions are no longer dominated by the rest of the samples, leading to estimators no longer treating those data points as outliers and fitting them more closely, ultimately resulting in less biased solutions, decreased misspecification, and more accurate ITE estimates. The following section describes our proposed method in detail.

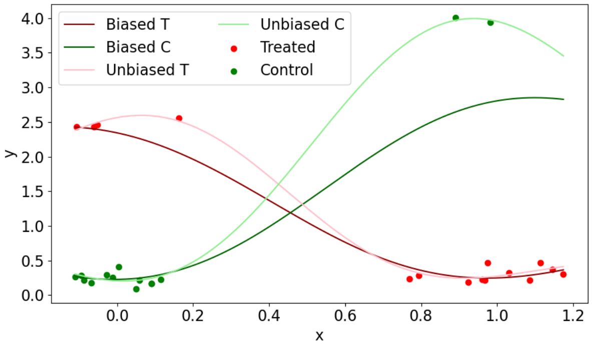

FIGURE 1. An example highlighting model misspecification issue. T = Treated, C = Control. The difference in ITE error is almost twice as in ATE.

IV. DEBIASING GENERATIVE TREES

As described in the previous section, model misspecification can be caused by underrepresented or missing data regions. Reweighing partially addresses this problem, but struggles with ITE estimation, not to mention propensity score approximators are subject to misspecification too. To avoid these pitfalls, we tackle misspecification through undersmoothness by augmenting the original data with new data points that carry useful information and help achieve the final estimators better ITE predictions. As the injected samples are expected to be informative to the learners, the overall data complexity increases as a consequence. Moreover, because this is a data augmentation procedure, it is estimator agnostic, that is, it can be used by any existing estimation methods. It is also worth pointing out that simply modelling and oversampling the entire joint distribution would not work as the learnt joint would include any existing data imbalances. In other words, underrepresented data regions would remain in minority, not addressing the problem at hand.

This observation led us to a conclusion that there is a need to identify smaller data regions, or clusters, and model their distributions in separation instead, giving us control over which areas to sample from and with what ratios. To achieve this, we incorporate recently proposed Generative Trees [11], which retain all the benefits of standard decision trees, such as simplicity, speed and transparency. They can also be easily extended to ensembles of trees, often improving the performance significantly. In practice, a standard decision tree regressor is used to learn the data. Once the tree is constructed, the samples can be assigned to tree leaves according to the learnt decision paths, forming distinct subpopulations that we are after. The distributions of these clusters are then separately modelled through Gaussian Mixture Models (GMMs). Similarly to decision trees, we again prioritise simplicity and ease of use

Algorithm 1 Debiasing Generative Trees

```
Input: $X$ - data set, E - estimator
Parameter: N - number of generated samples
Output: $E_D$ - debiased estimator
1: Let $X_G = \varnothing$.
2: Split $X$ into treated and control units ($X_T$ and $X_C$).
3: Train a Decision Tree regressor on $X_T$.
4: Map $X_T$ to tree leaves. Obtain subpopulations $S$.
5: Let $N_G = N/(2 \times len(S))$.
6: for $S_i$ in $S$ do
7:     Model $S_i$ with Gaussian Mixture Models. Obtain $G_i$.
8:     Draw $N_G$ samples from $G_i$. Store them in $X_G$.
9: end for
10: Repeat steps 3-9 for $X_C$.
11: Merge $X$ and $X_G$ into a single data set $X_M$.
12: Train estimator $E$ on $X_M$. Get debiased estimator $E_D$.
13: return debiased estimator $E_D$
```

here, which is certainly the case with GMMs. The next step is to sample equally from modelled distributions, that is, to draw the same amount of new samples per each GMM. In this way, we reduce data imbalances. A merge of new and original data is then provided to a downstream estimator, resulting in a less biased final estimator. Through experimentation, we find that splitting the original data at the beginning of the process into treated and control units and learning two separate trees for each group helps achieve better overall effect. A step-by-step description of the proposed procedure is presented in Algorithm 1.

As ensembles of trees almost always improve over simple ones, we incorporate Extremely Randomised Trees [62] for an additional performance gain. The procedure remains the same on a high level, differing only in randomly selecting inner trees at the time of sampling. Overall, we call this approach Debiasing Generative Trees (DeGeTs) as a general framework, with DeGe Decision Trees (DeGeDTs) and DeGe Forests (DeGeFs) for realisations with Decision Trees and Extremely Randomised Trees respectively.

There are a few important parameters to take care of when using the method. Firstly, depth of trees controls the granularity of identified subpopulations. Smaller clusters may translate to less accurate modelled distributions, whereas too shallow trees will bring the modelling closer to the entire joint that may result in not solving the problem of interest at all. The other tunable knob is the amount of new data samples to generate, where more data usually equates to a stronger effect, but also higher noise levels, which must be controlled to avoid destroying meaningful information in the original data. Finally, the number of components in GMMs is worth considering, where more complex distributions may require higher numbers of components.

As our method encourages higher modelling complexity, it is important to consider overfitting, which can be taken care of through the standard practice of tuning the above-mentioned hyperparameters and cross-validation. This can be done by using a downstream estimator's performance as a feedback signal as to which parameters work the best, which can also be tailored to a specific estimator of choice. The number of GMM components can be alternatively optimised through Bayesian Information Criterion (BIC) score. In order to make this method as general and easy to use as possible, we instead provide a set of reasonable defaults that we find work well across different data sets and settings. Default parameters: $max\_depth = \lceil \log_2 N_f \rceil - 1$, where N_f denotes the number of input features, $n\_samples = 0.5 \times size(training\_data)$, $n\_components \in [1, 5]$ — pick the one with the lowest BIC score.

In addition, we observe the fact that *DeGeTs* framework goes beyond applied Generative Trees and GMMs. This is because the data splitting part can, in fact, be performed by other methods, such as clustering. Consequently, GMMs can be substituted by any other generative models.

V. EXPERIMENTS

We follow recent literature (e.g. [50], [51], [52]) in terms of incorporated data sets and evaluation metrics. We start with defining the latter as different data sets use different sets of metrics. The source code that allows for a full replication of the presented experiments is available online[1] and is based on the *CATE benchmark*.[2]

There are a few aspects we aim to investigate. Firstly, how the established reweighing methods perform in individual treatment effect estimation. Secondly, how the choice of model class impacts estimation accuracy (misspecification). Thirdly, how our proposed method affects the performance of the base learners, and how it compares to other methods. Finally, we also study how our method influences the number of rules in pruned decision trees as an indirect measure of data complexity.

Although we do perform hyperparameter search to some extent in order to get reasonable results, it is not our goal to achieve the best results possible, hence the parameters used here are likely not optimal and can be improved upon more extensive search. The main reason is the setups presented as part of this work are intended to be as general as possible. This is why in our analysis we specifically focus on the relative difference in performance between settings rather than comparing them to absolute state-of-the-art results.

A. EVALUATION METRICS

The main focus of utilised metrics is on the quantification of the errors made by provided predictions. Thus, the metrics are usually denoted as ϵ_X, with error ϵ made with respect to prediction type X (lower is better). In terms of treatment outcomes, $\mathcal{Y}_t^{(i)}$ and $\hat{y}_t^{(i)}$ denote true and predicted outcomes respectively for treatment t and individual i. Thus, following the definition of ITE in (1), the difference $\mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}$ gives

[1] https://github.com/misoc-mml/undersmoothing-data-augmentation

[2] https://github.com/misoc-mml/cate-benchmark

a true effect, whereas $\hat{y}_1^{(i)} - \hat{y}_0^{(i)}$ a predicted one. Following this, we can define Precision in Estimation of Heterogeneous Effect (PEHE), which is the root mean squared error between predicted and true effects, as given in (5).

$$\epsilon_{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_1^{(i)} - \hat{y}_0^{(i)} - (\mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}))^2}. \quad (5)$$

Following the definition of ATE in (2), we measure the error on predicted ATE as the absolute difference between predicted and true average effects, formally written as in (6). Note, instead of expected values used in (2), here we switch to sample averages as the data sets used in experiments are of finite size, denoted in (6) with n.

$$\epsilon_{ATE} = \left| \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_1^{(i)} - \hat{y}_0^{(i)}) - \frac{1}{n}\sum_{i=1}^{n}(\mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}) \right|. \quad (6)$$

Given a set of treated subjects T that are part of sample E coming from an experimental study, and a set of control group C, we can define the true Average Treatment effect on the Treated (ATT) as per (7). It is a difference between the average outcome of the treated units and the average outcome of control units that come from experimental data. Note that $|A|$ denotes the cardinality of a set A and $A \cap B$ is the intersection of sets A and B.

$$\text{ATT} = \frac{1}{|T|}\sum_{i \in T}\mathcal{Y}^{(i)} - \frac{1}{|C \cap E|}\sum_{i \in C \cap E}\mathcal{Y}^{(i)}. \quad (7)$$

The error on predicted ATT is then defined as the absolute difference between the true and predicted ATT as in (8).

$$\epsilon_{ATT} = \left| \text{ATT} - \frac{1}{|T|}\sum_{i \in T}(\hat{y}_1^{(i)} - \hat{y}_0^{(i)}) \right|. \quad (8)$$

To measure the risk (or regret) of a policy (or treatment assignment) recommendation, let us define policy π that depends on background features x such that $\pi(x) = 1$ if $\hat{y}_1 - \hat{y}_0 > 0$; $\pi(x) = 0$ otherwise. Following such policy means recommending the treatment to any individual who will benefit (positive effect) from it based on predicted outcomes. The risk of applying such policy is defined as policy risk $\mathcal{R}_{pol}$ in (9).

$$\begin{aligned}\mathcal{R}_{pol} = 1 - (&\mathbb{E}[\mathcal{Y}_1|\pi(x) = 1]\,\mathcal{P}(\pi(x) = 1) \\ &+ \mathbb{E}[\mathcal{Y}_0|\pi(x) = 0]\,\mathcal{P}(\pi(x) = 0)), \quad (9)\end{aligned}$$

with mathematical expectation $\mathbb{E}[.]$ being switched to sample averages on data sets of finite size.

B. DATA

The baseline data sets we incorporate in our experiments are well-established and commonly used in the causal inference literature (see e.g. [28], [50], [51], [52]) to evaluate and compare estimation performances. We provide brief descriptions for each of the data sets; see respective references for additional details. These are also summarised in Table 1 and openly accessible online [63].

It is worth noting that real-world causal inference observational data sets are naturally "broken" as they inherently suffer from selection biases and covariate shifts, often in the form of distributional differences between treated and untreated units. Thus, in order to test causal estimators in conditions similar to real-world situations, a common practice is to purposefully "break" existing data sets by introducing biases and shifts. Our experiments incorporate such data sets as overcoming said data challenges, and measuring the degree of success via appropriate evaluation metrics, is the goal of this work.

TABLE 1. Summary of incorporated data sets.

data set	# samples (t/c)[a]	# features	outcome
IHDP	747 (139/608)	25	continuous
JOBS	3,212 (297/2,915)	17	binary
NEWS	5,000 (2,289/2,711)	3,477	continuous
TWINS	11,984 (5,992/5,992)	194	binary

[a] t = treated, c = control.

IHDP: Introduced by [64], based on Infant Health Development Program (IHDP) clinical trial [65]. The experiment measured various aspects of premature infants and their mothers, and how receiving specialised childcare affected the cognitive test score of the infants later on. We use a semi-synthetic version of this data set, where the outcomes are simulated through the NPCI package[3] (setting 'A') based on real pre-treatment covariates. Moreover, the treatment groups are made imbalanced by removing a subset of the treated individuals. We report errors on estimated PEHE and ATE (ϵ_{PEHE} in (5) and ϵ_{ATE} in (6) respectively) averaged over 1,000 realisations and split the data with 90/10 training/test ratios.

JOBS: This data set, proposed by [66], is a combination of the experiment done by [67] as part of the National Supported Work Program (NSWP) and observational data from the Panel Study of Income Dynamics (PSID) [68]. Overall, the data captures people's basic characteristics, whether they received a job training from NSWP (treatment), and their employment status (outcome). Here, we report ϵ_{ATT} (see (8)) and $\mathcal{R}_{pol}$ (see (9)) averaged over 10 runs with 80/20 training/test ratio splits.

NEWS: Introduced by [50], which consists of news articles in the form of word counts with respect to a predefined vocabulary. The treatment is represented as the device type (mobile or desktop) used to view the article, whereas the simulated outcome is defined as the user's experience. Similarly to IHDP, we report ϵ_{PEHE} (see (5)) and ϵ_{ATE} (see (6)) errors for this data set, averaging over 50 realisations with 90/10 training/test ratio splits.

TWINS: The data set comes from official records of twin births in the US in years 1989-1991 [69]. The data are preprocessed to include only individuals of the same sex and where each of them weight less than 2,000 grams.

[3] https://github.com/vdorie/npci

The treatment is represented as whether the individual is the heavier one of the twins, whereas the outcome is the mortality within the first year of life. As both factual and counterfactual outcomes are known from the official records, that is, mortality of both twins, one of the twins is intentionally hidden to simulate an observational setting. Here, we incorporate the approach taken by [28], where new binary features are created and flipped at random (0.33 probability) in order to hide confounding information. We report ϵ_{ATE} (as in (6)) and ϵ_{PEHE} (as in (5)) for this data set, averaged over 10 iterations with 80/20 training/test ratio splits.

C. SETUP

We incorporate the following estimators.

Base Learners: Linear methods: Lasso (l1) and Ridge (l2). Simple Trees: pruned Decision Trees, Extremely Randomised Trees (ET) [62]. Boosted Trees: CatBoost [70], LightGBM [71]. Kernel Ridge regression with nonlinearities. Dummy regressor returning the mean as a reference only.

Reweighing Methods: Causal Forest [5], Double Machine Learning (DML) [1], and Meta-Learners [4] in the form of T and X variations.

Debiasing Generative Trees: Our proposed method. We include the stronger performing *DeGeF* variation.

A general approach throughout all conducted experiments was to train a method on the training set and evaluate it against appropriate metrics on the test set. 5 base learners were trained and evaluated in that way: l1, l2, Simple Trees, Boosted Trees and Kernel Ridge. DML and Meta-Learners were combined with different base learners as they need them to solve intermediate regression and classification tasks internally. This resulted in $3 \times 5 = 15$ combinations of distinct estimators. Similarly, *DeGeF* was combined with the same 5 base learners to investigate how they react to our data augmentation method. Causal Forest and dummy regressor were treated as standalone methods. Overall, we obtained 27 distinct estimators per each data set. In terms of Simple and Boosted Trees, we defaulted to ETs and CatBoost respectively. For NEWS, due to its high-dimensionality, we switched to computationally less expensive Decision Trees and LightGBM instead.

As our *DeGeF* method is a data augmentation approach, it affects only the training set that is later used by base learners. It does not change the test set in any way as the test portion is used specifically for evaluation purposes to test how methods generalise to unseen data examples. More specifically, *DeGeF* injects new data samples to the existing training set, and that augmented training set is then provided to base learners.

Hyperparameter search was also performed wherever applicable, though not too extensive to keep our study as general and accessible as possible. The following is a list of base learners and their hyperparameters we explored. ETs: $max\_leaf\_nodes \in \{10, 20, 30, None\}$, $max\_depth \in \{5, 10, 20\}$. Kernel Ridge: $alpha \in \{0, 1e-1, 1e-2, 1e-3\}$, $gamma \in \{1e-2, 1e-1, 0, 1e+1, 1e+2\}$, $kernel \in \{rbf, poly\}$, $degree \in \{2, 3, 4\}$. CatBoost: $depth \in \{6, 8, 10\}$, $l2\_leaf\_reg \in \{1, 3, 10, 100\}$. LightGBM: $max\_depth \in \{5, 7, 10\}$, $reg\_lambda \in \{0, 0.1, 1, 5, 10\}$. Causal Forest: $max\_depth \in \{5, 10, 20\}$. For ETs, CatBoost, LightGBM and Causal Forest we set the number of inner estimators to 1000. To find the best set of hyperparameters, we performed 5-fold cross-validation. When it comes to *DeGeF*, we set the number of estimators to 10. The other parameters, like number of new samples, tree depth and GMM components, were set to defaults as recommended in the description of the framework. All randomisation seeds were set to a fixed number (1) throughout all experiments.

Most of our experimental runs were performed on a Linux based machine with 12 CPUs and 60 GBs of RAM. More demanding settings, such as NEWS combined with tree-based methods, were delegated to one with 96 CPUs and 500 GBs of RAM, though such a powerful machine is not required to complete those runs.

D. RESULTS

We incorporate the following estimator names throughout the presented tables: **l1** - Lasso, **l2** - Ridge, **kr** - Kernel Ridge, **dt** - Decision Tree, **et** - Extremely Randomised Trees, **cb** - CatBoost, **lgbm** - LightGBM, **cf** - Causal Forest, **dml** - Double Machine Learning, **xl** - X-Learner, **degef** - our *DeGeF* method. Combinations of the methods are denoted with a hyphen, for instance, 'dml-l1'. All presented numbers (excluding relative percentages explained below) denote means and 95% confidence intervals.

Estimation performance. Tables 2 - 5 present the main results, where we specifically focus on: a) relevant to a given data set metrics, and b) changes in performance relative to a particular base learner. The latter is calculated as $((r_a - r_b)/r_b) \times 100\%$, where r_a and r_b denote results of advanced methods and base learners respectively. The reason for analysing these relative changes rather than absolute values is because in this study we are specifically interested in how more complex approaches (including ours) affect the performance of the base learners, even if not reaching state-of-the-art results. For example, if a relative change for *xl-et* reads '−20', it means this estimator decreased the error by 20% when compared to plain *et* learner for that particular metric. Changes greater than zero denote an increase in errors (lower is better).

Data Complexity: Table 6 shows the number of rules obtained from a pruned Decision Tree while trained on original data and augmented by *degef*. The purpose of this experiment is to explain the mechanism through which our method affects estimation performance of downstream learners. Here, we interpret the number of induced tree rules as a model complexity level required to fit the data accurately. Since our goal is bias reduction and undersmoothing, an increase in model complexity after data augmentation would be desirable. Thus, by measuring model complexity

TABLE 2. Results for IHDP data set.

name	ϵ_{ATE}	Δ%	ϵ_{PEHE}	Δ%[a]
dummy	4.408 ± .103	-	7.898 ± .473	-
cf	0.397 ± .045	-	3.387 ± .318	-
l1	0.981 ± .106	-	5.790 ± .514	-
dml-l1	0.387 ± .043	−60.52	7.782 ± .691	34.42
tl-l1	0.273 ± .033	−72.19	7.858 ± .678	35.73
xl-l1	0.282 ± .034	−71.27	7.660 ± .678	32.31
degef-l1	1.051 ± .107	7.15	5.809 ± .514	0.33
l2	0.974 ± .104	-	5.786 ± .514	-
dml-l2	0.381 ± .040	−60.91	7.859 ± .691	35.82
tl-l2	0.273 ± .034	−72.02	7.810 ± .679	34.99
xl-l2	0.287 ± .034	−70.53	7.723 ± .678	33.47
degef-l2	1.093 ± .107	12.16	5.820 ± .514	0.58
dt	0.636 ± .084	-	4.025 ± .402	-
dml-dt	1.262 ± .116	98.50	6.679 ± .570	65.95
tl-dt	0.406 ± .044	−36.22	8.012 ± .698	99.07
xl-dt	0.529 ± .065	−16.81	7.317 ± .653	81.79
degef-dt	0.542 ± .075	−14.83	3.882 ± .384	−3.55
kr	0.356 ± .031	-	2.276 ± .170	-
dml-kr	0.616 ± .059	73.06	8.174 ± .728	259.16
tl-kr	0.167 ± .010	−53.02	8.024 ± .706	252.60
xl-kr	0.247 ± .023	−30.65	7.847 ± .698	244.82
degef-kr	0.316 ± .031	−11.18	2.149 ± .181	−5.58
et	0.519 ± .074	-	3.093 ± .322	-
dml-et	0.869 ± .082	67.61	6.532 ± .563	111.23
tl-et	0.306 ± .042	−41.01	7.445 ± .643	140.75
xl-et	0.453 ± .053	−12.63	6.875 ± .597	122.32
degef-et	0.394 ± .052	−24.03	2.818 ± .273	−8.89
cb	0.404 ± .038	-	2.179 ± .210	-
dml-cb	1.123 ± .052	177.88	6.976 ± .580	220.18
tl-cb	0.224 ± .027	−44.48	7.715 ± .664	254.10
xl-cb	0.388 ± .044	−3.97	6.894 ± .604	216.42
degef-cb	0.328 ± .032	−18.73	2.013 ± .190	−7.63
lgbm	0.412 ± .052	-	2.866 ± .273	-
dml-lgbm	1.516 ± .142	268.30	7.544 ± .632	163.25
tl-lgbm	0.255 ± .028	−38.10	8.002 ± .678	179.25
xl-lgbm	0.435 ± .046	5.53	7.602 ± .650	165.29
degef-lgbm	0.397 ± .051	−3.54	2.691 ± .250	−6.09

Metrics are *mean* ± 95%CI (lower is better).
[a] Δ% = change over the baseline (negative means improvement).

TABLE 3. Results for JOBS data set.

name	ϵ_{ATT}	Δ%	$\mathcal{R}_{pol}$	Δ%[a]
dummy	0.029 ± .000	-	0.326 ± .000	-
cf	0.025 ± .000	-	0.294 ± .000	-
l1	0.005 ± .000	-	0.296 ± .000	-
dml-l1	0.012 ± .000	146.75	0.366 ± .000	23.43
tl-l1	0.012 ± .000	140.15	0.374 ± .000	26.10
xl-l1	0.022 ± .000	361.49	0.356 ± .000	20.16
degef-l1	0.054 ± .012	1010.26	0.296 ± .000	0.00
l2	0.034 ± .000	-	0.296 ± .000	-
dml-l2	0.008 ± .000	−77.14	0.374 ± .000	26.20
tl-l2	0.007 ± .000	−79.25	0.370 ± .000	24.75
xl-l2	0.011 ± .000	−67.37	0.361 ± .000	21.91
degef-l2	0.056 ± .009	62.76	0.296 ± .000	0.00
dt	0.029 ± .000	-	0.365 ± .000	-
dml-dt	0.149 ± .000	408.57	0.336 ± .000	−7.80
tl-dt	0.035 ± .000	21.07	0.351 ± .000	−3.68
xl-dt	0.037 ± .000	27.98	0.296 ± .000	−18.75
degef-dt	0.048 ± .014	64.01	0.335 ± .015	−8.12
kr	0.017 ± .000	-	0.400 ± .000	-
dml-kr	0.007 ± .000	−61.39	0.374 ± .000	−6.52
tl-kr	0.005 ± .000	−70.54	0.305 ± .000	−23.81
xl-kr	0.003 ± .000	−80.58	0.279 ± .000	−30.32
degef-kr	0.019 ± .012	11.82	0.299 ± .013	−25.16
et	0.006 ± .000	-	0.276 ± .000	-
dml-et	0.099 ± .000	1686.11	0.353 ± .000	27.66
tl-et	0.010 ± .000	86.50	0.295 ± .000	6.81
xl-et	0.004 ± .000	−36.17	0.235 ± .000	−14.87
degef-et	0.015 ± .009	167.98	0.270 ± .014	−2.24
cb	0.026 ± .000	-	0.308 ± .000	-
dml-cb	0.010 ± .000	−60.23	0.368 ± .000	19.42
tl-cb	0.026 ± .000	−0.50	0.250 ± .000	−18.86
xl-cb	0.045 ± .000	72.93	0.239 ± .000	−22.56
degef-cb	0.019 ± .007	−26.61	0.257 ± .030	−16.51
lgbm	0.029 ± .000	-	0.247 ± .000	-
dml-lgbm	0.191 ± .000	555.20	0.387 ± .000	56.81
tl-lgbm	0.004 ± .000	−86.33	0.305 ± .000	23.62
xl-lgbm	0.021 ± .000	−29.31	0.297 ± .000	20.20
degef-lgbm	0.021 ± .007	−27.62	0.283 ± .024	14.83

Metrics are *mean* ± 95%CI (lower is better).
[a] Δ% = change over the baseline (negative means improvement).

this way we can inspect the existence and strength of such desirable properties. Note that sensitivity to noise and overfitting are not the subjects of interest in this particular experiment.

VI. DISCUSSION

In terms of IHDP data set (Table 2), the classic methods (*dml*, *tl*, and *xl*) strongly improve in ATE, but can also be unstable as it is the case with *dml*, specifically *dml-cb* and *dml-lgbm*. Against PEHE, the situation is much worse as those methods significantly decrease in performance when compared to the base learners, not to mention catastrophic setbacks in the worst cases (deltas above 200%). Note that not a single traditional method improves in PEHE (all deltas positive). Our *degef*, on the other hand, often improves in both ATE and PEHE (see negative deltas). Even in the worst cases with *l1* and *l2*, *degef* is still very stable and does not destroy the predictions as it happened with the other approaches. Thus, our method clearly offers the best improvements in PEHE and competitive predictions in ATE while providing a good amount of stability.

In the JOBS data set (Table 3), classic methods again achieve strong improvements in average effect estimation (ATT) in best cases, though they can be substantially worse as well (e.g. *dml-et*). In policy predictions, an equivalent of ITE, traditional techniques are even less likely to provide improvements, except the *X-Learner*. With respect to *degef*, it can also worsen the quality of predictions in ATT, as shown with *degef-l1*, though it does not get as bad as with *dml-et*. However, even in that worst example, policy predictions are not destroyed. The best cases in *degef*, on the other hand, achieve strong improvements in policy. Similarly to IHDP, here *degef* provided solid improvements in ITE predictions (policy), while staying on par with traditional methods in ATT, obtaining reasonable improvements and keeping the worst cases still better than the worst ones in the other methods, proving again its stability.

TABLE 4. Results for TWINS data set.

name	ϵ_{ATE}	Δ%	ϵ_{PEHE}	Δ%[a]
dummy	0.033 ± .002	-	0.318 ± .004	-
cf	0.064 ± .001	-	0.323 ± .005	-
l1	0.042 ± .000	-	0.319 ± .004	-
dml-l1	0.028 ± .003	−33.55	0.318 ± .004	−0.29
tl-l1	0.052 ± .001	23.80	0.324 ± .005	1.59
xl-l1	0.053 ± .001	25.46	0.322 ± .004	0.71
degef-l1	0.064 ± .004	53.10	0.323 ± .004	1.18
l2	0.047 ± .002	-	0.320 ± .004	-
dml-l2	0.042 ± .001	−11.32	0.334 ± .009	4.25
tl-l2	0.042 ± .000	−10.47	0.337 ± .011	5.19
xl-l2	0.042 ± .001	−10.95	0.335 ± .010	4.83
degef-l2	0.067 ± .004	41.28	0.324 ± .004	1.10
dt	0.004 ± .005	-	0.319 ± .004	-
dml-dt	0.070 ± .011	1859.14	0.327 ± .002	2.53
tl-dt	0.062 ± .000	1631.81	0.334 ± .004	4.67
xl-dt	0.059 ± .000	1549.54	0.323 ± .004	1.20
degef-dt	0.064 ± .013	1697.62	0.349 ± .005	9.37
kr	0.045 ± .001	-	0.320 ± .004	-
dml-kr	0.055 ± .028	20.87	0.323 ± .012	0.99
tl-kr	0.050 ± .000	9.18	0.334 ± .006	4.45
xl-kr	0.043 ± .002	−4.96	0.325 ± .007	1.73
degef-kr	0.033 ± .004	−27.17	0.320 ± .004	0.15
et	0.027 ± .006	-	0.322 ± .003	-
dml-et	0.047 ± .002	74.36	0.320 ± .005	−0.32
tl-et	0.051 ± .000	87.25	0.327 ± .006	1.76
xl-et	0.050 ± .001	85.14	0.323 ± .006	0.53
degef-et	0.054 ± .007	96.91	0.335 ± .002	4.23
cb	0.039 ± .000	-	0.319 ± .004	-
dml-cb	0.078 ± .011	99.66	0.328 ± .002	2.65
tl-cb	0.051 ± .000	31.77	0.331 ± .008	3.65
xl-cb	0.048 ± .002	22.63	0.323 ± .006	1.04
degef-cb	0.051 ± .003	31.48	0.326 ± .004	2.06
lgbm	0.038 ± .000	-	0.327 ± .005	-
dml-lgbm	0.034 ± .007	−10.56	0.362 ± .008	10.90
tl-lgbm	0.042 ± .002	9.79	0.393 ± .009	20.34
xl-lgbm	0.039 ± .002	2.02	0.366 ± .009	12.18
degef-lgbm	0.042 ± .002	8.61	0.328 ± .006	0.56

Metrics are *mean* ± 95%CI (lower is better).
[a] Δ% = change over the baseline (negative means improvement).

TABLE 5. Results for NEWS data set.

name	ϵ_{ATE}	Δ%	ϵ_{PEHE}	Δ%[a]
dummy	2.714 ± .212	-	4.381 ± .361	-
cf	0.544 ± .089	-	3.907 ± .481	-
l1	0.244 ± .068	-	3.370 ± .365	-
dml-l1	0.233 ± .062	−4.50	2.469 ± .269	−26.73
tl-l1	0.298 ± .052	22.13	2.166 ± .201	−35.74
xl-l1	0.220 ± .045	−9.75	2.152 ± .186	−36.14
degef-l1	0.225 ± .048	−7.86	3.370 ± .361	0.00
l2	0.260 ± .068	-	3.371 ± .366	-
dml-l2	0.236 ± .080	−9.08	5.108 ± .394	51.52
tl-l2	0.173 ± .030	−33.33	4.182 ± .343	24.06
xl-l2	0.174 ± .036	−33.09	4.162 ± .345	23.45
degef-l2	0.178 ± .041	−31.64	3.366 ± .362	−0.16
dt	0.344 ± .076	-	2.717 ± .277	-
dml-dt	4.523 ± .783	1216.23	5.875 ± .676	116.18
tl-dt	0.329 ± .062	−4.12	2.638 ± .222	−2.92
xl-dt	0.290 ± .060	−15.47	2.639 ± .263	−2.87
degef-dt	0.355 ± .080	3.22	2.727 ± .266	0.35
kr	0.715 ± .133	-	3.316 ± .367	-
dml-kr	2.544 ± .256	255.79	4.186 ± .399	26.25
tl-kr	0.198 ± .150	−72.27	2.677 ± .290	−19.26
xl-kr	0.229 ± .112	−68.00	2.695 ± .297	−18.72
degef-kr	0.582 ± .102	−18.61	3.256 ± .349	−1.80
et	0.276 ± .051	-	2.063 ± .200	-
dml-et	x	-	x	-
tl-et	x	-	x	-
xl-et	x	-	x	-
degef-et	0.290 ± .052	5.13	2.013 ± .167	−2.40
cb	0.127 ± .029	-	1.880 ± .179	-
dml-cb	x	-	x	-
tl-cb	x	-	x	-
xl-cb	x	-	x	-
degef-cb	x	-	x	-
lgbm	0.162 ± .045	-	2.074 ± .241	-
dml-lgbm	1.461 ± .181	799.12	3.240 ± .386	56.27
tl-lgbm	0.161 ± .033	−0.81	1.861 ± .138	−10.25
xl-lgbm	0.131 ± .042	−19.41	2.005 ± .228	−3.31
degef-lgbm	0.151 ± .042	−6.78	2.038 ± .228	−1.71

Metrics are *mean* ± 95%CI (lower is better).
Estimators marked with ‘x’ – no results due to excessive training time.
[a] Δ% = change over the baseline (negative means improvement).

TWINS data set (Table 4), proved to be very difficult for all considered methods when it comes to PEHE, though they did not worsen the predictions as well. Some good improvements in ATE can be observed, but also noticeable decreases in performance in the worst cases (combinations with *dt*). Our method behaves similarly to the classic ones, offering occasional gains and keeping the decreases in reasonable bounds. The stability of *degef* is especially noticeable in PEHE as the worst decrease (*degef-dt*) is still better than in other methods.

The last data set, NEWS (Table 5), showed the traditional approaches can provide some improvements in PEHE as well, at least in their best efforts, though performance decreases are also noticeable in the worst ones. They also offer quite stable improvements in ATE, except extremely poor *dml-dt*. The *X-Learner* performs particularly well across both metrics (most deltas negative). Our proposed method offers reasonable gains in ATE as well, while keeping performance decreases at bay even in the worst efforts. Even though *degef* provides little improvement in PEHE, it does not destroy individualised predictions either. Overall, this data set showcases superior stability properties of *degef* particularly well, making it a preferable choice if small but safe performance gains are desirable over potentially higher but riskier improvements.

In general terms, the results show that performance can vary substantially depending on the model class, even within the same advanced method (*dml, xl, degef*). For instance, *DML* proved to work particularly well with *L1* and *L2* as base learners, whereas *X-Learner* often outperforms *T-Learner*, adding more stability to the results as well. Our proposed technique usually offers significant improvements in ITE predictions in best cases, often better than traditional methods, while keeping the predictions stable even in the worst examples. Classic methods are clearly strong in ATE estimates, but can struggle in individualised predictions. Overall, these methods (*dml, xl*) proved to be less stable than ours, where the worst cases can perform quite poorly, especially *dml*. This makes *degef* a safer choice on average when considering various estimators, even more so when achieving the best possible performance is not considered a priority.

The observation that the choice of model class can significantly impact estimation performance opens up a more general question about possible reasons behind said performance differences. We investigated this question closely from the perspective of hyperparameters in another study [72], which offers rather surprising lessons. We have found that hyperparameter selection plays a significant role in this in the sense that, if done optimally (with access to a tuning oracle), the performance differences among individual estimators become small, rendering model selection secondary and suggesting that *free lunches* are possible under the right conditions. A wider implication of this is that causal estimators are generally comparable with respect to *potential* performance (only potential performance because optimal tuning is impossible in practice) and that some of the performance differences we found can be attributed to imperfect model evaluation, which in turn suffers from the same challenges as causal estimation itself – missing counterfactuals and covariate shifts. As a result, while theoretically many estimators are capable of similar performance levels, the best one can do in practice is to perform hyperparameter tuning as thoroughly as possible (which we do in our experiments here) to reduce the influence of model evaluation imperfections.

We also investigate the number of rules in pruned Decision Trees as a proxy for data complexity and required model complexity to accurately fit the data. As presented in Table 6, *degef* significantly increases the amount of rules across all data sets, translating to an increase in data complexity. This proves that augmented data encourages richer model families that are more likely to include the true DGP, subsequently leading to reduced bias, undersmoothing, and decreased misspecification. In addition, we observe that modest data complexity increases in IHDP and JOBS correlate with strong *degef* gains in ITE estimation in those two data sets, whereas a much bigger difference in TWINS (from 9.6 to 59.1) correlated with considerably lower prediction performance gains (Table 4). This suggests there is a practical limit to increased data complexity beyond which performance benefits decrease.

After combining all the results together, we can observe that *degef*: a) improves effect predictions (Tables 2 - 5), and b) increases data complexity (Table 6). Both points essentially demonstrate the positive effects our method has on prediction performance (point (a)) and specific mechanisms enabling such benefits (point (b)). More specifically, *degef* encourages higher modelling complexity (undersmoothing) through increased complexity of the augmented data (point (b)). This reduces bias and misspecification, which in practical terms improves prediction performance, even under covariate shift (as per point (a)). In terms of theoretical guarantees, we rely on [7] and [9], which provide a thorough formal analysis of the problem of model misspecification and undersmoothing respectively.

TABLE 6. Number of rules in a pruned decision tree.

data set	original data	augmented data	Δ%
IHDP	33.6 ± 2.0	53.3 ± 2.6	58.63
JOBS	6.0 ± 0.0	11.3 ± 5.3	88.33
TWINS	9.6 ± 0.9	59.1 ± 11.9	515.63
NEWS	19.4 ± 2.5	32.0 ± 4.7	64.95

Numbers are *mean* ± 95%CI.
We interpret number of tree rules as a proxy for model complexity required to fit the data (more rules, more complex model), which is an indirect proof for increased data complexity as a result of *degef* data augmentation (Δ% column), confirming the desired undersmoothing effect and reduced model misspecification have been achieved successfully.

VII. LIMITATIONS

In terms of possible limitations of our method, we assume the data sets we work with have relatively low noise levels. This is because in noisy environments, the inner GMMs would likely pick up a lot of noise and thus sampling from them would result in even more noisy data samples. The result would be the opposite of what we aim for, that is, to increase data complexity and bring new informative samples, not to introduce bias in the form of noise. Thus, our method would likely worsen base learners' performances in such environments. Furthermore, we expect extremely high-dimensional data sets may cause computational issues due to the increasing depth of the inner trees. This is partly why setting a reasonable depth limit is important. Our proposed method is also subject to the standard set of assumptions (SUTVA and strong ignorability; see Section II-B). Thus, scenarios that violate those are outside the applicability of the method.

VIII. CONCLUSION

In this work, we proposed Debiasing Generative Trees (DeGeTs), a novel data augmentation method based on generative trees for improved estimation of heterogeneous causal effects. Data augmented by *DeGeTs* through oversampling underrepresented data regions reduces bias and undersmooths causal estimators trained on the data. Higher modelling complexity of downstream learners achieved this way enriches the model family that is more likely to include the true DGP and hence reduces model misspecification. This in practice results in more accurate predictions, even with covariate shift, especially in individualised estimation where the consequences of misspecification are exacerbated.

Our key finding is that our proposed approach offers significantly better performance improvements in individual effect estimation as compared to traditional reweighing procedures while staying competitive on average effect tasks. Our method also exhibits better stability in terms of provided gains than other approaches, rendering it a safer option overall. Furthermore, we show through our experiments that the choice of model class can significantly affect achieved performance, and that reweighing methods can struggle on individualised estimation tasks. This links with our recent research on hyperparameters suggesting that tuning alone can be a source of major differences between performances [72]. Note, however, that hyperparameter optimisation is highly non-trivial in causal settings as it suffers from the same challenges as causal estimation itself.

In terms of possible future directions, it might be interesting to investigate the feasibility of replacing generative trees with neural networks to handle extremely high-dimensional problems. Another direction would be to instantiate *DeGeTs* framework with alternative methods, such as standard clustering and generative neural networks. Furthermore, extending our approach to data sets with high degree of noise could increase its applicability to a wider set of real-world tasks. In addition, an in-depth theoretical analysis of specific mechanisms behind increased estimation robustness of the proposed method may further explain its effectiveness.

REFERENCES

[1] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *Econ. J.*, vol. 21, no. 1, pp. 1–68, Feb. 2018.

[2] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 846–866, Sep. 1994.

[3] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, Apr. 1983.

[4] S. R. Kunzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 10, pp. 4156–4165, Mar. 2019.

[5] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *Ann. Statist.*, vol. 47, no. 2, pp. 1148–1178, Apr. 2019.

[6] J. Wen, C.-N. Yu, and R. Greiner, "Robust learning under uncertain test distributions: Relating covariate shift to model misspecification," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 631–639.

[7] H. White, "Consequences and detection of misspecified nonlinear regression models," *J. Amer. Stat. Assoc.*, vol. 76, no. 374, pp. 419–433, Jun. 1981.

[8] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey, "Double machine learning for treatment and causal parameters," Centre Microdata Methods Pract. (CEMMAP), London, U.K., Working Paper CWP49/16, Sep. 2016, doi: 10.1920/wp.cem.2016.4916.

[9] W. Newey, F. Hsieh, and J. Robins, "Undersmoothing and bias corrected functional estimation," Dept. Econ., Massachusetts Inst. Technol. (MIT), Cambridge, MA, USA, Working Paper 98-17, Oct. 1998.

[10] C. Hansen, D. Kozbur, and S. Misra, "Targeted undersmoothing: Sensitivity analysis for sparse estimators," *Rev. Econ. Statist.*, vol. 105, no. 1, pp. 101–112, Jan. 2023.

[11] A. Correia, R. Peharz, and C. P. de Campos, "Joints in random forests," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33. New York, NY, USA: Curran Associates, 2020, pp. 11404–11415.

[12] R. Balestriero, I. Misra, and Y. LeCun, "A data-augmentation is worth a thousand samples: Analytical moments and sampling-free training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 19631–19644.

[13] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*.

[14] Y. Hou and M. Navarro-Cía, "A computationally-inexpensive strategy in CT image data augmentation for robust deep learning classification in the early stages of an outbreak," *Biomed. Phys. Eng. Exp.*, vol. 9, no. 5, Sep. 2023, Art. no. 055003.

[15] A. Kiran and S. S. Kumar, "A comparative analysis of GAN and VAE based synthetic data generators for high dimensional, imbalanced tabular data," in *Proc. 2nd Int. Conf. Innov. Technol. (INOCON)*, Mar. 2023, pp. 1–6.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[17] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1322–1328.

[18] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Appl. Soft Comput.*, vol. 83, Oct. 2019, Art. no. 105662.

[19] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Inf. Sci.*, vol. 512, pp. 1214–1233, Feb. 2020.

[20] M. Ilse, J. M. Tomczak, and P. Forré, "Selecting data augmentation for simulating interventions," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 4555–4562.

[21] A. Balashankar, X. Wang, Y. Qin, B. Packer, N. Thain, E. Chi, J. Chen, and A. Beutel, "Improving classifier robustness through active generative counterfactual data augmentation," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 127–139.

[22] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Mach. Learn. Appl.*, vol. 9, Sep. 2022, Art. no. 100375.

[23] S. Pitis, E. Creager, A. Mandlekar, and A. Garg, "MoCoDA: Model-based counterfactual data augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 18143–18156.

[24] F. A. Tan, D. Hazarika, S.-K. Ng, S. Poria, and R. Zimmermann, "Causal augmentation for causal sentence classification," in *Proc. 1st Workshop Causal Inference NLP*, 2021, pp. 1–20.

[25] S. C. M. Gowda, S. Joshi, H. Zhang, and M. Ghassemi, "Pulling up by the causal bootstraps: Causal data augmentation for pre-training debiasing," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manag.* New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 606–616.

[26] S. Athey, G. Imbens, J. Metzger, and E. Munro, "Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations," 2019, *arXiv:1909.02210*.

[27] B. Neal, C.-W. Huang, and S. Raghupathi, "RealCause: Realistic causal inference benchmarking," 2020, *arXiv:2011.15007*.

[28] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[29] Y. Hu, Y. Wu, L. Zhang, and X. Wu, "A generative adversarial framework for bounding confounded causal effects," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 13, pp. 12104–12112.

[30] K. Kuang, Y. Li, B. Li, P. Cui, H. Yang, J. Tao, and F. Wu, "Continuous treatment effect estimation via generative adversarial de-confounding," *Data Mining Knowl. Discovery*, vol. 35, no. 6, pp. 2467–2497, Nov. 2021.

[31] M. El Bouchattaoui, M. Tami, B. Lepetit, and P.-H. Cournède, "Causal dynamic variational autoencoder for counterfactual regression in longitudinal data," 2023, *arXiv:2310.10559*.

[32] I. Bica, J. Jordon, and M. van der Schaar, "Estimating the effects of continuous-valued interventions using generative adversarial networks," 2020, *arXiv:2002.12326*.

[33] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[34] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[35] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 1–37, Jul. 2020.

[36] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," 2020, *arXiv:2002.02770*.

[37] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educ. Psychol.*, vol. 66, no. 5, pp. 688–701, Oct. 1974.

[38] S. Bickel, M. Bruckner, and T. Scheffer, "Discriminative learning for differing training and test distributions," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 81–88.

[39] T. T. Cai, H. Namkoong, and S. Yadlowsky, "Diagnosing model performance under distribution shift," 2023, *arXiv:2303.02011*.

[40] H. Zhang, H. Singh, M. Ghassemi, and S. Joshi, "'Why did the model fail?': Attributing model performance changes to distribution shifts," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, Jul. 2023, pp. 41550–41578.

[41] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," 2021, *arXiv:2108.13624*.

[42] K. Imai and M. Ratkovic, "Covariate balancing propensity score," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 76, no. 1, pp. 243–263, Jan. 2014.

[43] B. K. Lee, J. Lessler, and E. A. Stuart, "Weight trimming and propensity score weighting," *PLoS ONE*, vol. 6, no. 3, Mar. 2011, Art. no. e18174.

[44] R. Gutman, E. Karavani, and Y. Shimoni, "Propensity score models are better when post-calibrated," 2022, *arXiv:2211.01221*.

[45] X. Nie and S. Wager, "Quasi-oracle estimation of heterogeneous treatment effects," *Biometrika*, vol. 108, no. 2, pp. 299–319, May 2021.

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

[47] M. J. van der Laan and D. Rubin, "Targeted maximum likelihood learning," *Int. J. Biostatistics*, vol. 2, no. 1, pp. 1–38, Jan. 2006.

[48] M. J. van der Laan and S. Rose, "Why machine learning cannot ignore maximum likelihood estimation," in *Handbook of Matching and Weighting Adjustments for Causal Inference*. Boca Raton, FL, USA: CRC Press, 2023, pp. 483–500.

[49] S. Samothrakis, A. Matran-Fernandez, U. Abdullahi, M. Fairbank, and M. Fasli, "Grokking-like effects in counterfactual inference," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[50] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, 2016, pp. 3020–3029.

[51] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Montréal, QC, Canada: Curran Associates, Dec. 2018, pp. 2638–2648.

[52] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3076–3085.

[53] C. Shi, D. Blei, and V. Veitch, "Adapting neural networks for the estimation of treatment effects," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. New York, NY, USA: Curran Associates, 2019, pp. 1–11.

[54] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?" in *Proc. 36th Int. Conf. Mach. Learn.*, May 2019, pp. 872–881.

[55] J. Yoon, J. Jordon, and M. van der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," in *Proc. Int. Conf. Learn. Represent.*, Feb. 2018, pp. 1–22.

[56] T. Vanderschueren, J. Berrevoets, and W. Verbeke, "NOFLITE: Learning to predict individual treatment effect distributions," *Trans. Mach. Learn. Res.*, pp. 1–17, Jul. 2023.

[57] A. Alaa, Z. Ahmad, and M. van der Laan, "Conformal meta-learners for predictive inference of individual treatment effects," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 1–22.

[58] Y. Zhang, C. Shi, and S. Luo, "Conformal off-policy prediction," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2023, pp. 2751–2768.

[59] M. Oprescu, J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, and U. Shalit, "B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, Jul. 2023, pp. 26599–26618.

[60] T. Vanderschueren, A. Curth, W. Verbeke, and M. Van Der Schaar, "Accounting for informative sampling when learning to forecast treatment outcomes over time," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, Jul. 2023, pp. 34855–34874.

[61] K. E. Rudolph, N. T. Williams, C. H. Miles, J. Antonelli, and I. Diaz, "All models are wrong, but which are useful? Comparing parametric and nonparametric estimation of causal effects in finite samples," *J. Causal Inference*, vol. 11, no. 1, Nov. 2023, Art. no. 20230022.

[62] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[63] D. Machlanski, "Treatment effect estimation benchmarks," *IEEE Dataport*, Aug. 2023, doi: 10.21227/0v4q-nn37.

[64] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *J. Comput. Graph. Statist.*, vol. 20, no. 1, pp. 217–240, Jan. 2011.

[65] J. Brooks-Gunn, F. R. Liaw, and P. K. Klebanov, "Effects of early intervention on cognitive function of low birth weight preterm infants," *J. Pediatrics*, vol. 120, no. 3, pp. 350–359, Mar. 1992.

[66] J. A. Smith and P. E. Todd, "Does matching overcome LaLonde's critique of nonexperimental estimators?" *J. Econ.*, vol. 125, nos. 1–2, pp. 305–353, Mar. 2005.

[67] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *Amer. Econ. Rev.*, vol. 76, no. 4, pp. 604–620, 1986.

[68] R. H. Dehejia and S. Wahba, "Propensity score-matching methods for nonexperimental causal studies," *Rev. Econ. Statist.*, vol. 84, no. 1, pp. 151–161, Feb. 2002.

[69] D. Almond, K. Y. Chay, and D. S. Lee, "The costs of low birth weight," *Quart. J. Econ.*, vol. 120, no. 3, pp. 1031–1083, Aug. 2005.

[70] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31. New York, NY, USA: Curran Associates, 2018, pp. 1–11.

[71] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Adv. Neural Inf. Process. Syst.*, vol. 30. New York, NY, USA: Curran Associates, 2017, pp. 1–9.

[72] D. Machlanski, S. Samothrakis, and P. Clarke, "Hyperparameter tuning and model evaluation in causal effect estimation," 2023, *arXiv:2303.01412*.

DAMIAN MACHLANSKI received the B.Eng. degree in computer science from West Pomeranian University of Technology, Szczecin, Poland, and the M.Sc. degree in artificial intelligence from the University of Essex, U.K., where he is currently pursuing the Ph.D. degree in computer science. His research interests include machine learning and causality, with a particular focus on the methods of treatment effect estimation and causal discovery.

SPYRIDON SAMOTHRAKIS received the Ph.D. degree in computer science in 2014, with a focus on optimal game playing in multi-player games. He is currently a Senior Lecturer and the Deputy Director of the Institute for Analytics and Data Science, University of Essex. He has been involved extensively with businesses. His current research interests include meta-learning and reinforcement learning.

PAUL CLARKE is currently a Professor of social statistics with the Institute for Social and Economic Research, University of Essex. His work involves developing statistical methods for applications in the social sciences, with a special focus on methods for the causal and longitudinal analysis of survey data. His current research interests include the interface between statistical methods and machine learning.