Results

## Results

### Proof of concept

|  |
| --- |
| **Figure** 1: Bias on effect estimates for different ML algorithms in three different simulated causal simulations (a and b). Sample sizes are so large that stochastic effects can be excluded (1000 observations and 500 repetitions). Effects of the ML models were inferred using average conditional effects. Row a) shows results for simulations with uncorrelated predictors with effect sizes (=1.0, =0.0, and =1.0). Row b) shows the results for simulations with X1 and X2 being strongly correlated (Pearson correlation factor = 0.9) but only X1 affects y. |

### Data-poor simulation

|  |
| --- |
| **Figure** 2: Bias and variance of estimated effects in data-poor situations. N = 50, 100, and 600 observations of 100 weakly correlated predictors were simulated. True effects in the data generating model were =1.0, =0.0, and the other 98 effects were equally spaced between 0 and 1. Models were fitted to the simulated data (1000 replicates) with the optimal hyperparameters (except for LM, which doesn’t have hyperparameters). Hyperparameters were selected based on the minimum MSE of () (green) or the prediction error (based on ) (red). Bias and variance were calculated for and . Effects for ) were approximated using ACE. |

#### Hyper-parameter sensitivity analysis

|  |
| --- |
| **Figure** 3: Results of hyperparameter tuning for Neural Networks (NN), Boosted Regression Trees (BRT), Random Forests (RF), and Elastic Net (EN) for 100 observations with 100 predictors. The influence of the hyperparameters on effect () (bias, variance, and MSE), and the predictions of the model, (), (bias, variance, and MSE) were estimated by a multivariate generalized additive model (GAM). Categorical hyperparameters (activation function in NN) were estimated as fixed effects. The responses (bias, variance, MSE) were centered so that the categorical hyperparameters correspond to the intercepts. The variable importance of the hyperparameters was estimated by a random forest with the MSE of the effect (first plot) or the prediction (second plot) as the response. Red dots correspond to the best predicted set of hyperparameters (based on a random forest), in the first plot for the minimum MSE of the effect for and in the second plot for the minimum MSE of the predictions . |

### Case Study

|  |
| --- |
| **Figure** 4: Difference between causal and conventional ML models for in-distribution and out-of-distribution predictions. In a simulated setting, the task is to predict Crop yield based on Plant growth (data-generating model is shown in the figure). Climate is an unobservable confounder and has effects on Plant growth and Pest (growth). In the first scenario, i.e. in-distribution predictions, Climate did not change, i.e. patients were exposed to the same climatic conditions; here the difference in predictive performance for the model with and without Pest growth is marginal (predictive performance was measured by R2). In the second theoretical setting, the climatic conditions changed (the effects of Climate on Plant growth and Pest are now zero). Using the previously trained models, the model without Pest deficit performed significantly worse than the model with Pest (plot with out-of-distribution predictions). |