Causal Inference for Personalized Treatment Effect Estimation for given Machine Learning Models

Johannes Rust German Research Center for Artificial Intelligence GmbH
Enrique-Schmidt-Str. 5
28359 Bremen, Germany
Email: johannes.rust@dfki.de Serge Autexier German Research Center for Artificial Intelligence GmbH
Enrique-Schmidt-Str. 5
28359 Bremen, Germany
Email: serge.autexier@dfki.de

***Abstract*—We propose a causal machine learning inference pipeline that combines a given predictive machine learning model with analytical estimations of average treatment effects. It enables to utilize any predictive model for causal inference, which makes it easy to adapt the approach to existing systems. By first estimating the average treatment effect of an intervention on predictors instead of the outcome variable, a causal relationship between an intervention and a wide range of variables is determined. Next, artificial samples are created that are evaluated using the given predictive model to link interventions and outcomes and also allows inferring measurements of uncertainty. Finally, simulations using again the given predictive model are performed to compute measurements of confidence and that allow to compare – according to the given predictive model – the effect of specific treatments. We furthermore demonstrate how this inference engine can be adapted to a privacy-preserving federated learning environment where training data is horizontally distributed across multiple datasets without compromising on our approach's accuracy. The approach has been evaluated on a use case with a predictive model for the quality of life score of cancer patients, to determine medical interventions to improve their individual quality of life score.**

I. Introduction

With machine learning (ML) providing strong prediction capabilities for a wide variety of tasks, it is a promising tool to use in many domains, e.g. economics or medicine. Machine learning models are typically only used to predict a target variable value or class. However, especially in the medical domain it is desirable to use machine learning to assess the outcome of a specific treatment on the target variable value, which is not an information readily represented in the data. The estimation of an outcome depending on an intervention is known as *causal inference*.

From detecting diseases from blood values, analyzing medical images for cancer or risk estimations, many techniques have been developed in recent years [1]. It comes apparent that introducing machine learning in such safety critical environments yields multiple challenges. For one, wrong predictions can result in irreversible damage and harm to patients. This necessitates the development of trustworthy systems. For a machine learning model to be trustworthy, it needs transparency in its decision-making process and clearly communicate a measurement of confidence for its predictions. Hence, there is a need that causal inference methods [16] are transparent and provide a confidence.

Another effort is the distribution of input data that is needed to train reliable models. Suitable training data is often scattered over multiple healthcare facilities and cannot be aggregated directly because of legal reasons like data privacy policies. Therefore, a flexible system is needed, that can be trained in a federated learning environment, while still maintaining the transparency needed to earn the users trust.

We propose an approach that can fit a single machine learning model to be used for causal inference. After reviewing the state of the art in section II we first present some background information on the research project that provides the context of this work. In section IV we elaborate how our method can be used to evaluate multiple treatments or interventions and also present how it can be adapted to a federated learning and homomorphic encryption. The approach is evaluated on two scenarios and the results presented in section V. In section VII we also elaborate how the decision-making process is easy to interpret by users, leading to better explainable machine learning and discuss in section VIII how the approach is used in our project to evaluate medical interventions that improve the Quality of Life of breast- and prostate cancer patients. Section IX concludes the paper and presents future work.

II. Related Work

A. Causal Inference

Causal Inference [16] was proposed to create interpretable, robust but also capable machine learning models. Its central approach is to measure causal relationships. Most AI models only link common patterns and distributions in the input data to the output data. The direction of causality is ignored and providing explanations for prediction does not mitigate this shortcoming. Measuring the average treatment effect (ATE) is a common way to measure causality [2], [3]. A commonly used method are meta learners [10] which are families of ML models trained on disjoint subsets of the training data containing the different treatments and thus estimate the ATE directly. They have the advantage of being able to detect complex patterns and also calculate heterogeneous treatment effects, meaning that the estimated treatment effect is not

constant, but moderated by other influences (e.g a patient's gender) – but require retraining for the different training data subsets. These causal models also do not solve the problem of limited explainability. The causality provides a direction of dependencies, but a meta learner is still as uninterpretable as a normal ML model. Also, to adapt this approach to a federated learning environment, the meta learners must be trained in a federated way as well, yielding numerous challenges.

B. Explainable Artificial Intelligence

Explainable Artificial Intelligence is an artificial intelligence system like machine learning that has special attributes to make its output comprehensible for the user by providing additional information on how the result was inferred. Methods for explainability can be categorized in post-hoc methods and ante-hoc methods. Post-hoc methods describe the methods that aim to make a machine learning model or, more generally, a black box model interpretable or describing its outputs after the model itself has already been created. [12] Common approaches are surrogate models, which are interpretable by design and mimic the models input/output behavior. Surrogate models can also be fitted to explain only few or even a single prediction. Ribeiro et al. proposed local surrogates that approximate a linear model around a sample to estimate feature attribution models [11]. Lundberg et al. [7] proposed a way to estimate Shapley-values by adapting local surrogate models to approximate a model's response to data points near the sample in the feature space and also added improvements for specific model types [8], [9]. Some researchers argued that post-hoc explainability is not enough to produce sufficient transparency and that explainability must be considered in the early development space of an AI system. Using interpretable model architectures such as linear regression or decision trees is an easy approach, but their predictive capabilities are limited.

III. Background

The work presented in the paper has been conducted in the context of research project ASCAPE aiming to provide a platform that gives predictions for future Quality of Life (QoL) issues for breast and cancer patients. It shall support medical staff to find suitable medical interventions that address current QoL issues and avoid future ones. A set of different machine learning architectures and approaches are used to predict the QoL scores and the risk of specific issues that influence the patients QoL. Local ML models are trained only on the patient data that is available in the respective healthcare sites. Global models are federated learned models trained on datasets from a cluster of healthcare sites. Also, global encrypted models are trained by collecting homomorphically encrypted data from all healthcare on the cloud and that can be used for predictions on the cloud for homomorphically encrypted patient data.

An obvious approach to let machine learning models propose interventions is to train dedicated models. They predict the probability that a certain intervention is chosen. They can be trained on existing patient data using performed treatments as labels. However, this approach is flawed. Since most ML models like neural networks are too complex to be interpreted by humans, the predictions that are being made by these models are not transparent. To provide more insight to the predictions, explainability methods like SHAP[7] could be used to provide feature attributions and provide a limited amount of explainability. However, even feature attribution values cannot be easily interpreted in this scenario. A certain value in a patients medical data having a high feature attribution for the chance that a certain treatment is proposed is still unintuitive information for the user. Moreover, using the treatments that have been proposed by the medical staff as ground truth for the AI models limits their capabilities. At best, they propose the same treatments the medical staff would have selected anyway, since they are optimized on their behavior.

IV. Method

In this work we use capital letters such as $X, Y, Z, \dots$ to denote variables, calligraphic letters $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ to denote sets of variables, lower-case letters $x, y, z, \dots$ to denote values of feature variables, the range/set of possible values of some feature variable X as $\mathrm{range}(X)$, and $\mathrm{range}(\mathcal{X})$ be the cartesian product of the possible values of all variables $X \in \mathcal{X}$ in lexicographic order.

Let D be dataset that contains a set of variables $C = \{\mathcal{X}, \mathcal{Z}, \mathcal{Y}\}$, with $\mathcal{X}$ being variables that are used as input variables, $\mathcal{Z}$ being a set of interventions or treatments and Y a variable that shall be predicted. We denote the dataset as a set of tuples $\vec{x}_t, \vec{x}_{t+1}, \vec{y}$ where $\vec{x}$ is a vector of all variables $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ at a time point t. $\vec{x}_{t+1}$ is the state of $\vec{x}$ at the following time point $t+1$ and $\vec{y}$ is the vector of all outcomes $Y \in \mathcal{Y}$ at time point $t+1$. Between time point t and $t+1$ interventions take place. These are denoted by $\vec{z}$ which is a vector of all interventions $Z \in \mathcal{Z}$. When splitting the dataset into cohorts, we denote $D_{\vec{z}}$ as a subset of D where the treatments in $\vec{x}$ are the same as $\vec{z}$:

$$D_{\vec{z}} = \{(\vec{x}_t, \vec{x}_{t+1}, \vec{y}) \in D | \vec{x} \downarrow_{\mathcal{Z}} = \vec{z}\} \tag{1}$$

The control cohort $D_{\vec{0}}$ consists of pairs in D where no intervention was performed:

$$D_{\vec{0}} = \{(\vec{x}_t, \vec{x}_{t+1}, \vec{y}) \in D | \vec{x} \downarrow_{\mathcal{Z}} = \vec{0}\} \tag{2}$$

Causal inference is usually based on the average treatment effect on some variable Y, which following [16] can be defined as

$$\begin{aligned} \forall Z \in \mathcal{Z} : ATE(Y, Z) &= \frac{\partial\, \mathbb{E}(Y|Z)}{\partial Z} \\ &= \mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0) \end{aligned} \tag{3}$$

assuming the treatments Z are binary, i.e. $\mathrm{range}(Z) = \{0, 1\}$.

If the ATE is dependent on sample values $\vec{x}$ for the variables in $\mathcal{X}$, it is defined as conditional average treatment effect (CATE):

$$\begin{aligned} &\forall \vec{x} \in \mathrm{range}(\mathcal{X}), Z \in \mathcal{Z} : \\ &\quad \textit{CATE}(Y, Z, \vec{x}) = \frac{\partial\, \mathbb{E}(Y|Z, \vec{x})}{\partial Z} \\ &\qquad\qquad = \mathbb{E}(Y|Z=1, \vec{x}) - \mathbb{E}(Y|Z=0, \vec{x}) \end{aligned} \tag{4}$$

To train models that can propose treatments that the medical staff might not have considered and support its predictions with sound explanations, the AI models must be able to assess how treatments influence the patient's health status.

Therefore, for our approach the prediction process is split into three steps. We assume that all variables X_{t+1} are dependent on their previous state X_t at time point t. We also assume that an outcome Y_t (and hence also its prediction) is dependent on X_t, but not directly dependent on its previous state X_{t-1}. An example for such an outcome is the Quality of Life, which depends mainly on the current situation rather than on the values of these at previous times.

First, the average treatment effect (ATE) for each medical intervention is determined for every variable X in the patient's health record. Then, the average treatment effect is used to predict the change of the patient's health record:

$$\forall Z \in \mathcal{Z} : ATE(X, Z) = \frac{\partial\, \mathbb{E}(X|Z)}{\partial Z} \tag{5}$$

$$= \mathbb{E}(X|Z = 1) - \mathbb{E}(X|Z = 0) \tag{6}$$

The ATE on X is then used to predict Y_{t+1} by estimating X_{t+1}, given the hypothesis that a certain medical intervention is being performed. Since Y_{t+1} is dependent on state X_{t+1} and X_{t+1} is dependent on X_t and Z, a causal estimation of Y_{t+1} can be done given X_t and Z. Formally, we define our approach as

$$\begin{aligned} \forall \vec{x} \in \mathrm{range}(\mathcal{X}), Z \in \mathcal{Z} : \mathrm{CATE}^*(Y, Z, \vec{x}) = \\ \mathbb{E}(Y|Z = 1, \vec{x} \oplus (\textit{ATE}(X, Z = 1)|X \in \mathcal{X})) \\ - \mathbb{E}(Y|Z = 0, \vec{x} \oplus (\textit{ATE}(X, Z = 0)|X \in \mathcal{X})) \end{aligned} \tag{7}$$

where $\oplus$ denotes sum of two vectors and $(\textit{ATE}(X, Z)|X \in \mathcal{X})$ the vector consisting of the values $\textit{ATE}(X, Z)$ for X taken from $\mathcal{X}$ in lexicographic order.

Now, let $m_Y : \mathrm{range}(\mathcal{X}) \times \mathcal{Z} \to Y$ be a machine learning model predicting/approximating $\mathbb{E}(Y|Z = z, x)$ for outcome Y based on an input vector $\vec{x} \in \mathrm{range}(\mathcal{X})$ and a treatment $z \in \mathcal{Z}$.

Replacing appropriately in the definition (7) of $\textit{CATE}^*$ the occurrence of *ATE* by f and $\mathbb{E}(Y|\dots)$ by m_Y we obtain the definition of our algorithm

$$\begin{aligned} \forall Y \in \mathcal{Y}, Z \in \mathcal{Z} : \textit{CATE}^+(Y, Z, D) = \\ m_Y(\vec{x} \oplus (ATE(X, D)_{Z=1}|X \in \mathcal{X}), Z = 1) \\ -m_Y(\vec{x} \oplus (ATE(X, D)_{Z=0}|X \in \mathcal{X}), Z = 0) \end{aligned} \tag{8}$$

Splitting up the inference pipeline into these steps yields more transparent prediction results and leverages the models to a causal inference pipeline.

The process to predict the personalized treatment effect for a sample (e.g. patient data) at hand consists of 3 steps: first, we estimate the average effect of treatments on the predictor variables (section IV-A) and use these in a second step to estimate the development of the sample (patient) after the treatment (section IV-B). Third, we use these to select the best treatment and assess there average effect as well as some upper and lower bounds (section IV-C).

A. Estimation of Average Treatment Effects

The average treatment effect is a common instrument to evaluate the causal relation between a treatment and an outcome. For this approach it is helpful to analyze the dataset like it is collected in a study. To calculate the ATE as reliably as possible, a randomized controlled study would be needed. Here, each participant is assigned to a test cohort or control cohort. The test cohort receives a treatment and the control cohort receives no treatment or, if applicable, a placebo. The difference of representative measurements can then be used to determine the ATE. However, this study design is not suitable outside trials where participants agree that they might be assigned to the test cohort and not receive a treatment. For patients in cancer-treatment, only an observational study design can be used since it does not interfere with the patient's treatment. Here, data can be collected or processed in two different ways: In longitudinal studies data is collected from a patient multiple times. Treatment effects can be approximated by comparing data before and after a treatment. In cross-sectional studies, data is only collected once. Treatment effects can therefore not be collected directly by comparing data points from the same patient. Instead, patients with similar properties must be matched in order to compare patients that received a treatment and those who did not. However, this approach is more prone to biases and therefore less accurate, especially when matching high dimensional data.

Since for our approach we aim to estimate the influence of a treatment Z on the variables X in $\mathcal{X}$, some additional conditions and assumptions must hold. Each feature variable X_{t+1} must be stochastically dependent on its predecessor X_t. We identified three cases of properties that X_t and X_{t+1} in relation to Z can have:

- Case 1: X_{t+1} is dependent on Z. An example could be physical fitness related to physical activity level.
- Case 2: X_{t+1} is independent of Z, but changes linearly to t: $X_{t+1} = X_t + c$ with c being constant for all t. An example is the age of the patient or the cancer type.
- Case 3: X_t is static for all t: $X_{t+1} = X_t$: $X_{t+1} = X_t$. This is equivalent to Case 2 with $c = 0$. Examples are past diseases or gender.

To make our approach as flexible as possible, we implemented solutions for both longitudinal datasets and cross-sectional datasets (using cohort matching). For the latter, cohort matching is done for every treatment $Z \in \mathcal{Z}$. The dataset is first split into two cohorts. Patients that received treatment Z are assigned to the test cohort, those who did not are assigned to the control cohort. However, simply comparing these two cohorts does only yield correlations, but not causalities. It cannot be determined if a patient is in a certain condition because of the treatment or if he received the treatment because of his condition. To mitigate this, patients with similar properties are paired based on their similarities of $X_{i,t}$. These variables should be static or not be dependent on the treatment (e.g. age at diagnosis, cancer type or gender). Before using cohort matching, all input variables should therefore be assessed using context knowledge and

assigned to one of the three cases we identified before. Only variables of case 3 are suitable to be used for cohort matching.

Two approaches were implemented and tested for cohort matching: Propensity score matching [4] and Mahalanobis distance matching [5]. While propensity score matching is widely used and easy to adapt on a wide range of data, it was criticized for not creating ideal results [6]. Therefore, Mahalanobis distance matching was also tested in our methods.

Since this approach is not suitable for high-dimensional input vectors that we have, we make use of our project's longitudinal study design and do not use cohort matching. For this, the dataset D is filtered into the test cohort $D_{\vec{z}}$ and the control cohort $D_{\vec{0}}$.

The expected change of a variable after a treatment can be estimated with

$$\forall \vec{x}_t, \vec{x}_{t+1} \in D_{\vec{z}} : \hat{x}_{t,Z=1} = \vec{x}_{t+1} - \vec{x}_t \tag{9}$$

To attribute for the fact that variables might change even if no intervention took place, the expected change without treatment is also calculated.

$$\forall \vec{x}_t, \vec{x}_{t+1} \in D_{\mathcal{Z}=0} : \hat{x}_{t,Z=0} = \vec{x}_{t+1} - \vec{x}_t \tag{10}$$

For the estimation of Z on Y the effects of $Z = 1$ on X and $Z = 0$ on X must be measured individually. We denote these values as $ATE_{Z=1}$ and $ATE_{Z=0}$.

$$ATE(X,D)_{Z=1} = \frac{1}{|D_{\vec{z}}|} \sum_{\vec{x}_t, \vec{x}_{t+1} \in D_{\vec{z}}} \hat{x}_{t,Z=1} \tag{11}$$

$$ATE(X,D)_{\mathcal{Z}=0} = \frac{1}{|D_{\mathcal{Z}=0}|} \sum_{\vec{x}_t, \vec{x}_{t+1} \in D_{\mathcal{Z}=0}} \hat{x}_{t,Z=0} \tag{12}$$

The overall average treatment effect can be calculated as the difference of $ATE_{Z=1}$ and $ATE_{Z=0}$.

$$\begin{aligned} \text{ATE}(X,D,Z) &= ATE(X,D)_{Z=1} - ATE(X,D)_{\mathcal{Z}=0} \quad (13) \\ &= \frac{1}{|D_{\vec{z}}|} \sum_{\vec{x}_t, \vec{x}_{t+1} \in D_{\vec{z}}} \hat{x}_{t,Z=1} \downarrow_X \\ &\quad - \frac{1}{|D_{\mathcal{Z}=0}|} \sum_{\vec{x}_t, \vec{x}_{t+1} \in D_{\vec{0}}} \hat{x}_{t,Z=0} \downarrow_X \quad (14) \end{aligned}$$

The above equation can handle all three cases of dependency we discussed earlier:

- For case 1, $\hat{x}_{t,Z=1} \downarrow_X \neq \hat{x}_{t,\mathcal{Z}=0} \downarrow_X$ and $ATE(X,D,Z)$ reflects the influence Z has on $\vec{x}_t \downarrow_X$
- For case 2, $\hat{x}_{t,Z=1} \downarrow_X = \hat{x}_{t,\mathcal{Z}=0} \downarrow_X + c$ since z has no influence on the difference between $\vec{x}_t \downarrow_X$ and $\vec{x}_t + 1$. The change that remains is only the independent term c.
- For case 3, $\hat{x}_{t,Z=1} \downarrow_X = \hat{x}_{t,\mathcal{Z}=0} \downarrow_X$ since z has no influence on the difference between $\vec{x}_t \downarrow_X$ and $\vec{x}_{t+1} \downarrow_X$

B. Creating Prospective Patient Samples

Since we calculated the average treatment effect for the input data and not the outcome directly, we can now estimate how the input data changes when a treatment is performed and how the input is expected to be at time point $t + 1$ when no treatment is performed. This estimated input is created by directly adding the ATEs to $\vec{x}$.

$$\vec{x}_{Z=1} = \vec{x} \oplus (ATE(X,D)_{Z=1} | X \in \mathcal{X}) \tag{15}$$

$$\vec{x}_{Z=0} = \vec{x} \oplus (ATE(X,D)_{Z=0} | X \in \mathcal{X}) \tag{16}$$

For each considered treatment in $\mathcal{Z}$, the input vector $\vec{x}_{Z=1}$ can be evaluated by making a prediction with $m_Y(\vec{x}_{Z=1}, Z = 1)$ (resp. $m_Y(\vec{x}_{Z=1,p}, Z = 1)$, $m_Y(\vec{x}_{Z=1,1-p}, Z = 1)$) which give respectively the prediction of the outcome of a treatment according to the average effect of the treatment Z and the outcomes with regard to the upper and the lower bounds of the treatment effect.

C. Personalized Treatment Effect Prediction

We use a machine learning model that can be a regression model or a classifier. This can be any model architecture and must not necessarily be trained on the same training samples that were used for the ATE inference. To identify the most promising treatment for an output variable Y, the treatment or treatment combination $\vec{z}$ that produces the highest predicted score is selected, which can be defined and simplified as follows since the value of $m_Y(\vec{x}, Z = 0)$ is always the same and can be factored out. Assuming the goal is to increase the value of some target variable Y (e.g. quality of life), the proposed treatment is determined by

$$\begin{aligned} &\text{argmax}_{Z \in \mathcal{Z}} \ m_Y(\vec{x}_{Z=1}, Z = 1) - m_Y(\vec{x}, Z = 0) \\ &= \text{argmax}_{Z \in \mathcal{Z}} \ m_Y(\vec{x}_{Z=1}, Z = 1) \end{aligned} \tag{17}$$

Based on this, we can predict the individual best treatment effect as follows:

$$\text{ITE}(Y, \vec{x}) = \text{CATE}^+(Y, \text{argmax}_{Z \in \mathcal{Z}} \ m_Y(\vec{x}_{Z=1}, Z = 1), \vec{x}) \tag{18}$$

If the output value shall be minimized, the lowest value is selected, i.e.

$$\text{argmin}_{Z \in \mathcal{Z}} \ m_Y(\vec{x}_{Z=1}, Z = 1) \tag{19}$$

Discussion. The proposed approach has been developed to assess the average effect of a single treatment and used it for personalized suggestions of the best treatments. However, the approach can also be used to assess the effect of combinations of treatments in an analogous way and thus supporting the selection of the best combinations of treatments. This is extremely valuable in the medical application domain of our project and, to our knowledge, extremely difficult if not infeasible due to low training data samples with classical methods.

D. Adaptation to Federated Learning

The approach was developed specifically to be used in a privacy-preserving environment. Here we assume we have a cluster of federated learning participants which we call edge nodes. Also, there is one instance that functions as cloud. The cloud coordinates the federated learning process and distributes information across the edge nodes. Each edge node communicates with the cloud, but never directly with other

edge node. Therefore, the system has a star topology. Any training data or samples used for predictions is located in one of the edge nodes and is never sent to the cloud nor is available for other edge nodes. It is assumed that the training data is horizontally distributed without duplicates on multiple edge nodes. Therefore, every edge node has a dataset with the same variables, but different samples. However, the approach in the following for horizontally distributed data should also be easily adaptable to vertically partitioned data.

To make knowledge about the average treatment effect available to all edge nodes, it is collected and aggregated in the cloud. Because averaging the individual treatment effects of the matched cohort pairs is linear, it is easy to combine them in the cloud with minimal loss compared to having all training stored in one instance.

We assume that we have the ATE of k edge nodes $\{1, ..., k\}$ with $ATE(X, Z)_l$ being the ATE of variable X after performing treatment Z, determined by the l-th edge node. Let n_l be the number of samples in a dataset of edge node l. We then define the overall combined ATEs called *global ATE (*ATE_{global}*)* of all edge nodes be their weighted average:

$$\mathrm{ATE}_{global}(X, Z) = \frac{1}{\sum_{l=1...k} n_l} \sum_{l=1...k} n_l \cdot \mathrm{ATE}_l(X, Z) \quad (20)$$

The global variance of the treatment effects is more difficult to estimate. Given that the mean value of the normal distribution is roughly the same for every edge node, the variance can be averaged as well:

$$\mathrm{VTE}_{global}(X, Z) = \frac{1}{\sum_{l=1...k} n_l} \sum_{l=1...k} n_l \cdot \mathrm{VTE}(X, Z)_l \quad (21)$$

Every time a dataset on an edge node is updated, the ATEs for that dataset are calculated again and stored locally as well as updated in the cloud. Once an ATE is updated, the global ATE is calculated by the cloud. In parallel to the updating of the ATEs, any predictive models might be updated as well. When a prediction is requested in one of the edge nodes, the global ATE is requested from the cloud and used to create hypothetical samples like described in section IV-B and used for predictions as described in section IV-C

V. Evaluation and Experiments

We evaluate our approach using three testing scenarios. The first scenario (ORB) is based on cancer patient dataset of our project. The second scenario uses simulated data from the ACIC 2016 challenge [14]. These two scenarios rely on cohort matching since the predictors are not available at multiple time points, i.e. the data is not sequential. For the third scenario we will use a synthetic dataset with sequential data. Here, all predictors, interventions and target variables are available for each time point. For the ORB dataset, there is no ground truth available, since each patient is only either part of the treatment group or the control group. We therefore only compare the ATE between estimating the ATE directly with Mahalanobis distance matching and averaging all CATE of our approach to get an overall ATE. For each dataset D, a model is trained based on all available predictors of that dataset. The effect of a treatment Z on a feature variable X (i.e. $f(X, Z = 1)$) is determined by Mahalanobis distance matching:

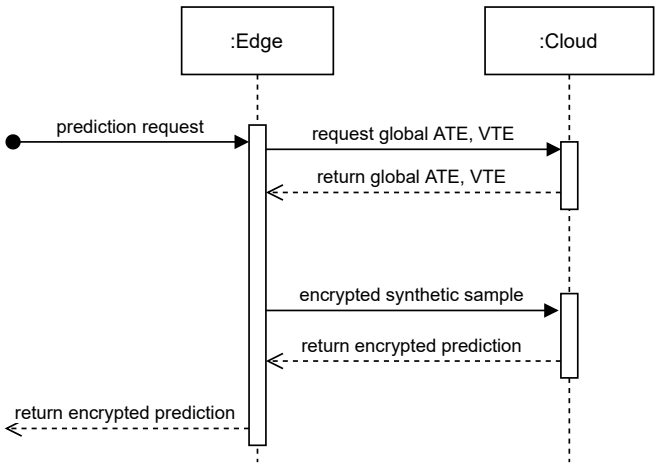

Figure 1. Sequence chart of prediction process with homomorphically encrypted models

Table I
Overview datasets created from the ORB dataset

Model name	Predictors until month	QoL at month	Number of samples
ORB-30-36	30	36	1138
ORB-30-60	30	60	1042
ORB-30-120	30	120	610
ORB-54-60	54	60	1042
ORB-54-120	54	120	610
ORB-108-120	108	120	610

$$\forall Z \in \mathcal{Z} : \mathrm{ATE}_{\text{classic}}(Y, Z) = \frac{1}{|P|} \sum_{\vec{y}_{Z=1}, y_{Z=0} \in P} y_{Z=1} - y_{Z=0} \quad (22)$$

$$\forall Z \in \mathcal{Z} : \mathrm{ATE}_{\text{new}}(Y, Z) = \frac{1}{|D|} \sum_{\vec{x} \in D} \mathrm{CATE}^{+}(Y, Z, \vec{x}) \quad (23)$$

a) ORB dataset: For this patient dataset Quality of Life of prostate cancer patients of the patient was collected using the LISAT-11 questionnaire [13] 36 months, 60 months and 120 months after the diagnosis. At baseline, 55 variables were collected. After that, every 6 months the lower urinary tract symptoms (LUTS), and the presence of bowel dysfunctions and erectile functions were reported by the patient. We created six datasets with the predictors and targets and trained respective models as show in Table I. Each model contains five treatments at baseline: Brachytherapy, Postoperative radiotherapy, External radiotherapy, Anti-androgen, GnRH-analogue and Hormonal therapy.

The results in Table II show that our approach computes results in the same order of magnitude as an established method and that it can handle multiple datasets at once.

VI. Synthetic Sequential Dataset

The synthetic sequential dataset mimics the structure of the original dataset of the ASCAPE project but is not privacy-sensitive. We create a dataset that contains n samples with

Table II
COMPARISON OF THE DETERMINED ATE BETWEEN OUR APPROACH ("NEW") AND USING COHORT MATCHING DIRECTLY ("CLASSIC")

	Brachytherapy		Ext. radiotherapy		Anti-androgen		GnRH-analogue		Hormonal therapy	
	new	classic	new	classic	new	classic	new	classic	new	classic
QoL 30-60m	1.59	**-0.37**	-3.42	**-3.37**	**-2.38**	-3.41	**-0.97**	-2.11	**-2.87**	-3.30
QoL 30-36m	**-0.15**	-1.35	**-2.68**	-3.52	**-1.92**	-2.92	**-1.19**	-1.67	**-2.20**	-3.13
QoL 30-120m	0.79	**-0.33**	**-2.54**	-6.69	**0.00**	-0.79	**-0.54**	-4.28	**-0.52**	-2.94
QoL 54-60m	**-0.03**	-0.12	**-3.18**	-3.78	**-1.87**	-3.14	**-0.80**	-1.37	**-1.92**	-2.95
QoL 54-120m	**1.62**	1.81	**-4.62**	-6.76	**0.13**	-1.30	**-0.57**	-4.19	**-0.27**	-2.62
QoL 108-120m	**1.42**	2.90	**-5.28**	-7.92	**-0.17**	-1.39	**-1.37**	-5.56	**-0.99**	-2,77

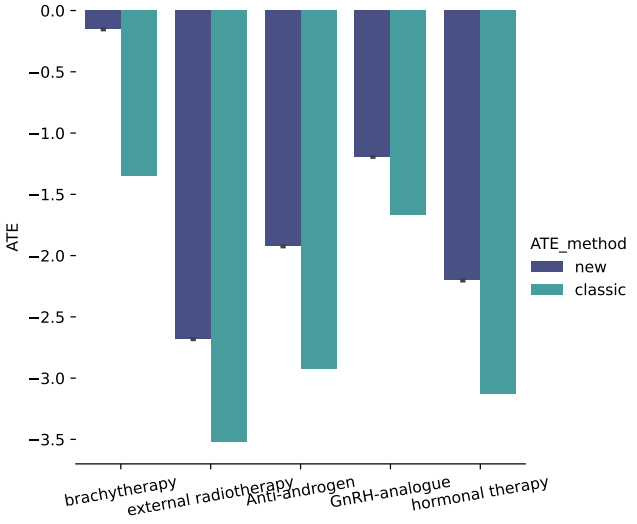

Figure 2. ATE for the ORB QoL 30-36m dataset determined by our approach ("new") compared to Mahalanobis distance matching ("classic")

$p = 196$ predictors, $i = 18$ interventions, and $l = 22$ target variables before and after an intervention took place. Let $x_t \sim \mathcal{N}(0,1)$ be a sample before an intervention took effect. Also, a matrix ATEs $\in \mathbb{R}^{p\times i}$ with random values from a uniform distribution ATEs $\sim \mathcal{U}^{p\times i}_{[0,1]}$ is generated that assigns an ATE for every intervention to every predictor. A random intervention $Z = z$ is drawn from the set of interventions and the predictors $\vec{x}_{t+1}$ are calculated as the sum of the predictors $\vec{x}_t$ and ATE(X, Z). Ground truth target values are inferred with a randomly initialized linear kernel $K \sim \mathcal{N}(0,1)^p$, simulating the stochastic relation between the predictors and the target.

$$y_{t+1} = x_{t+1} \cdot K \tag{24}$$

The predictors, interventions and target values are saved into a dataset with $n = 200$ samples. The ATEs and the kernel parameters remain unknown in the training process.

Models are then trained to predict the target variables based on the predictors of the same time point using the training pipeline and automatic model selection developed by Savić et al.[15]. Also, the $ATE(X, Z)$ is estimated for all $z \in Z$. For all interventions and all target variables (396 total results), the mean squared error of the individual treatment effects becomes 0, as well as a mean squared error of 0.0 for the targets after the interventions itself. This result is of course only achievable for a dataset where all variance in the labels is only dependent on the predictors. When adding a noise parameter ϵ to the dataset the mean average error for the ATE increases linearly to ϵ. However, it proves the capability of our approach that splitting causal inference is still solvable with machine learning. Both the ATEs and the kernel K were fully reconstructed in the training process.

VII. IMPROVE EXPLANATION OF TREATMENT PREDICTION EFFECTS

A key advantage of our approach is that it can also serve to improve explainability. In many cases of causal inference, the predictors are measurements of real, factual data, while the output is often more abstract, like risk estimations, customer satisfaction etc. The outcome is not directly dependent on the treatment, but rather on the predictors and only the predictors are stochastically dependent on the treatment. Predicting the outcome depending on a treatment is therefore legitimate, but by not assessing what the treatment directly changes, the inference is less intuitive for a user.

As an example, consider an elderly patient who fills out a QoL questionnaire and states that he has problems living a normal life due to limited physical abilities. Instead of using a model that proposes a nutritional consultation directly with a confidence score of 95%, the average treatment effect on the patient's health status is determined. It states that the patient's weight is expected to decrease by 8 kg when receiving nutritional consultation. A ML model predicting the score for a patient's mobility score predicts a significantly better score when the patient's weight is reduced by 8 kg. After consultation, the patient might state that he will try to lose at least 5 kg of weight. The doctor can examine using the ML model that even losing 5 kg is enough to expect a better QoL. This approach is more interactive and gives the users more information about why a certain treatment is proposed, increasing its interpretability and the user's trust in the system.

VIII. INTEGRATION IN A HEALTHCARE ENVIRONMENT

To use our approach in our cancer-patient related project, a preparation of the dataset D is required. First, the dataset needs to be partitioned into the subsets $\mathcal{X}$, $\mathcal{Z}$, and $\mathcal{Y}$. $\mathcal{X}$ can consist of any type of variables, i.e. nominal data, or numeric. Dates are made numeric by transforming them to days relative to a fixed date, e.g. date of birth. $\mathcal{Z}$ solely consists of binary variables. 1 indicates that a treatment was done and 0 that it was not. While most ML models support multiple outputs, we train a dedicated model for every $Y \in \mathcal{Y}$. This way, $\mathcal{Y}$ can be a mix of binary, nominal and numeric variables. Not all variables that resemble a treatment must be in $\mathcal{Z}$. Treatments that shall not be proposed can simply be assigned to $\mathcal{X}$ and be used as predictors.

IX. CONCLUSION

We presented a new approach to using average treatment effects to predict their effect on input variables and using a given predictive model like neural networks. We combined three desirable characteristics: We induce a causal relationship between a treatment and predictor variables, we provide an intermediate result that can easily be examined by users, and we can still make use of strong predictive capabilities

of the machine learning model. It was also shown that this approach can be adapted with little effort to other increasingly relevant applications in AI such as federated learning. The approach has been evaluated by applying it on available on data, where no ground truth on the treatment effect was available, but where it exhibits plausible results compared to the classical average treatment effect measurement method. A clear advantage of the presented approach is that it can be used with any available predictive model, needs no training of specific models and especially not only allows to consider the effect of a single treatment, but also the effect of the combination of several treatments.

We see further research potential in the development of a better approximation of the average treatment effects. The concept of meta learners providing heterogeneous treatment effect estimations potentially increases the accuracy and allows processing more complex data such as time-series. However, for usage in a federated environment, the meta learners would need to be trained in a federated manner as well, which might introduce further challenges.

ACKNOWLEDGMENT

This research was supported by the ASCAPE project. The ASCAPE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875351.

REFERENCES

[1] Wiens, Jenna, and Erica S. Shenoy. "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology." Clinical Infectious Diseases 66.1 (2018): 149-153.
[2] Angrist, Joshua, and Guido Imbens. "Identification and estimation of local average treatment effects." (1995).
[3] Jordà, Òscar, and Alan M. Taylor. "The time for austerity: estimating the average treatment effect of fiscal policy." The Economic Journal 126.590 (2016): 219-255.
[4] Rassen, Jeremy A., et al. "One-to-many propensity score matching in cohort studies." Pharmacoepidemiology and drug safety 21 (2012): 69-80.
[5] Rubin, Donald B. "Bias reduction using Mahalanobis-metric matching." Biometrics (1980): 293-298.
[6] King, Gary, and Richard Nielsen. "Why propensity scores should not be used for matching." Political Analysis 27.4 (2019): 435-454.
[7] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
[8] Lundberg, Scott M., et al. "From local explanations to global understanding with explainable AI for trees." Nature machine intelligence 2.1 (2020): 56-67.
[9] Lundberg, Scott M., et al. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery." Nature biomedical engineering 2.10 (2018): 749-760.
[10] Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156-4165.
[11] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)
[12] Holzinger, Andreas, et al. "Causability and explainability of artificial intelligence in medicine." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9.4 (2019): e1312.
[13] Post, Marcel WM, et al. "Predictors of health status and life satisfaction in spinal cord injury." Archives of physical medicine and rehabilitation 79.4 (1998): 395-401.
[14] Hill, Jennifer "2016 Atlantic Causal Inference Conference Competition" https://jenniferhill7.wixsite.com/acic-2016/competition (2016)
[15] Savić, Miloš, et al. "Analysis of Machine Learning Models Predicting Quality of Life for Cancer Patients." Proceedings of the 13th International Conference on Management of Digital EcoSystems. 2021.
[16] Pearl, Judea. "Causal inference." Causality: objectives and assessment (2010): 39-58.